ADVANCES IN PHYSICS

A QUARTERLY SUPPLEMENT of the PHILOSOPHICAL MAGAZINE

VOLUME 2 JULY 1953 NUMBER 7

PRICE per part 15*s.* 0*d.*
PRICE per annum £2 15*s.* 0*d.* post free

PRINTED AND PUBLISHED BY TAYLOR & FRANCIS LTD.
RED LION COURT, FLEET ST., LONDON E.C.4

CONTENTS

ADVANCES IN PHYSICS

A QUARTERLY SUPPLEMENT

of the

PHILOSOPHICAL MAGAZINE

VOLUME 2 JULY 1953 NUMBER 7

The Aurorae

By V. C. A. FERRARO
Queen Mary College, University of London

§ 1. INTRODUCTION

THE purpose of this article is to give an account of the Aurora polaris (also known as the Aurora Bolearis, or Northern Lights, in the northern hemisphere, and as the Aurora Australis, or Southern Lights in the southern hemisphere) and of the various theories which have been put forward to explain them.

The northern lights were known to the Greeks and Romans and the following striking account of Seneca (*Naturales Questiones*, I, 14, 15), quoted by Chapman and Bartels in *Geomagnetism* is of particular interest :

" Sometimes flames are seen in the sky, sometimes stationary or full of movements. Several kinds are known : the abysses, when beneath a luminous crown the heavenly fire is wanting, forming as it were the circular entrance to a cavern ; the tunns, when a great rounded flame in the form of a barrel is seen to move from place to place, or to burn immovable ; the gulfs, when the heavens seem to open and to vomit flames which before were hidden in its depths. There fires present the most varied colours : some are a vivid red ; others resemble a faint and dying flame ; some are white ; other scintillate ; others finally are of an even yellow . . . sometimes these fires are high enough to shine amongst the stars ; at others, so low that they might be taken for the reflection of a distant burning homestead or city. This is what happened under Tiberius, when the cohorts hurried to the succour of the colony of Ostia, believing it to be on fire."

The aurorae have always been a familiar phenomenon in northern countries and there are vivid descriptions of it by Norwegians as early as the thirteenth century. Later they were often confused with comets and in the sixteenth and seventeenth century were regarded as ill-portents. The aurora may last for only a few minutes but more often several hours or even the whole night. Extremely varied in form, its rapid changes in colour, intensity and position render it one of the most beautiful and striking of natural phenomena. Nevertheless, its total intensity is never very great and rarely exceeds that of the full moon.

The difficulty of photographing the aurora was not overcome till 1892, when Brendel used a combination of lens and plate which was fast enough to obtain exposures of only seven seconds. Later Störmer succeeded in photographing the aurora with exposures as short as half a second. Motion pictures have been obtained during magnetic storms by Vegard, Harang and W. Bauer and more recently colour pictures have been taken with 2 seconds exposure.

Many reports of aurorae are available (de Mairan 1773, Lovering 1868, Fritz 1873, Boller 1898, 1902), though largely scattered in meteorological and magnetic publications, and much valuable information was obtained during the polar years 1882–3 and 1932–3. Our knowledge of the position in space of the aurora is largely due to Störmer, Vegard and Krogness, *inter alia*.

There appears to be no essential difference between the northern and southern light; both are always accompanied by magnetic disturbances and at times of great magnetic storms the aurora can be seen in many parts of the world, even in very low latitudes.

I. GENERAL CHARACTERISTICS

§ 2. Classification of Auroral Forms

There are several types of auroral displays but most of them are composed of one or more principal forms, which can appear simultaneously. A special Committee of the International Geodetic and Geophysical Union has drawn up the following classification (Oslo, 1930) which includes two main types, namely (*a*) forms without ray structure and (*b*) those with ray structures.

(*a*) *Forms Without Ray Structures*

Homogeneous Quiet Arcs (Symbol HA)

These may have different breadth; when near the horizon they are bounded by a diffuse upper boundary and a sharp lower boundary which may be regular, like a low rainbow, or irregular and often stretching from horizon to horizon. Several parallel arcs may appear at the same time and merge together to form a large zone. The arcs may be incomplete, extending neither towards the east or west horizon.

Homogeneous Bands (Symbol HB)

These are less regularly shaped, more rapidly moving and may also have one or more folds. Their breadths vary from narrow bands to large bands resembling a curtain with sharp lower borders.

Pulsating Arcs (Symbol PA)

Whole arcs may light and disappear with a period of several seconds.

Diffuse Luminous Surface (Symbol DS)

A diffuse veil or glow over a great part of the sky, resembling clouds, with indistinct boundaries.

Pulsating Surface (Symbol PS)

Diffuse patches of light which appear and disappear at the same place and in the same irregular shape every few seconds.

Glow (Symbol G)

A feeble glow near the horizon of a white or reddish colour. It may often be the upper part of an arc, the lower border of which lies below the horizon.

(*b*) *Forms With Ray Structure*

These forms consist of short or long rays which seem to converge towards the magnetic zenith i.e. they follow the earth's magnetic lines of force.

Arc With Ray Structure (RA)

These resemble the homogeneous bands except that they consist of a series of rays which may be close together.

Draperies (D)

These consist of one or more bands with very long rays giving the appearance of a curtain.

Rays (R)

May be narrow or broad, short or long.

Corona (C)

This is often one of the most striking forms of the aurora and consists of rays radiating from the magnetic zenith and terminating around a central ring of light.

(*c*) *Flaming Aurora* (F)

This is a characteristically rapidly moving form, and consists of light moving rapidly up and down in the direction of the zenith. Photographs of some of these forms are shown in plates I–IV.

§ 3. Colour of the Aurora

The predominant colour is greenish yellow; sometimes it is bluish white, or red and occasionally deep red. According to Vegard, draperies are greenish yellow with reddish blue streamers and the lower border of draperies are sometimes dark red. According to Störmer, the colour of rays may change from bluish-green to reddish-violet.

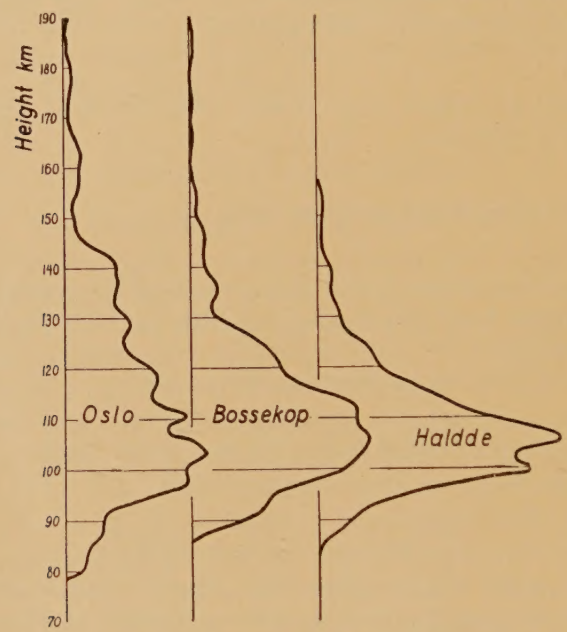
§ 4. The Height of the Aurora

The height of the aurora cannot be obtained from observations at a single station, but can be obtained from the azimuth and angular heights of identical points of the aurora photographed from two stations sufficiently far apart. At present regular height measurements are made from a network of auroral stations in Southern Norway under the direction of Störmer who considerably improved the methods of measurement in 1910 (Störmer 1911).

The method used is essentially that used in geodetic surveys, the ' base-line ' ranging from 26 km to 400 km. This method seems to have been first used by Mairan (1773) in 1726 by visual observations. Henry Cavendish (1790) using observations of an arc taken simultaneously at Cambridge and at Kimbolton, situated 23·8 miles north of it, deduced the height of the arc to be 112 km. His observations could also be interpreted to give a height of 84 km, but both limits agree well with modern measurements.

The long series of measurements taken by Störmer at Bossekop and by Vegard and Krogness at Haldde provide a great part of the information we have concerning the altitude of aurorae. Figure 1 shows the relative frequency of the lower limits of aurorae at different stations (after Störmer).

Fig. 1



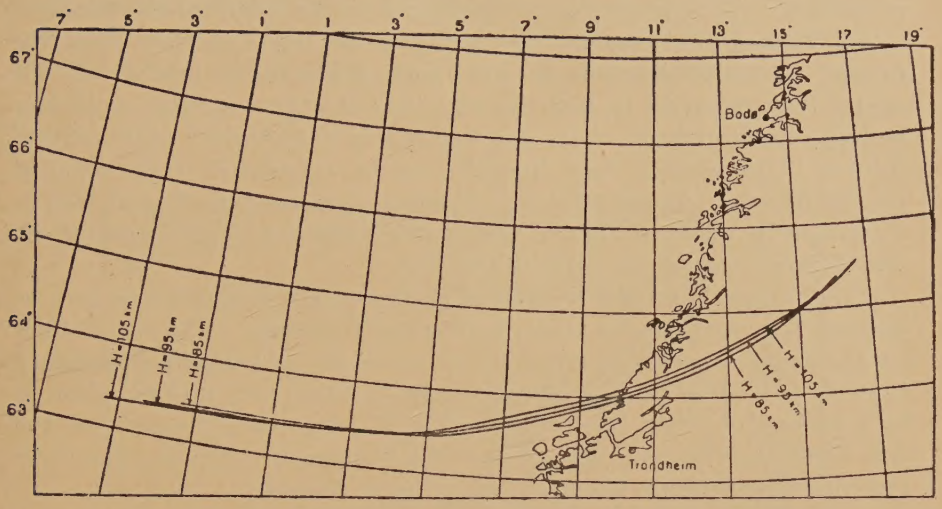
Diagrams showing the relative frequency of occurrence of the lower limits of aurorae at different heights, at three Norwegian stations (after Störmer).

All these curves exhibit maxima over a height interval extending between 106–107 km in northern and 103–104 km in southern Norway. Chapman however, casts doubt as to whether the undulations and double maximum in the curve for Haldde are really significant.

The heights of the lower limit of the various forms differ by only a few kilometers. Störmer considered it certain that draperies sometimes reach as low as 80 to 84 km whilst Harang and Bauer found a height of 65 km in March 1932 for the red lower border of a quiet arc (Harang and Bauer 1932). Occasionally on the other hand, the lower borders of draperies and arcs are seldom higher than 150 km, though they have been occasionally observed up to 200 km.

The rarity of low latitude aurorae makes their height difficult to determine but Götz estimates that the height of the great aurora of January 25, 1938, observed in Switzerland, must have been about 600 km.

Fig. 2



(*a*)

(*b*)

Geographical positions of the auroral arc of March 24, 1936 (*a*) and of the auroral curtain of January 24, 1936 (*b*) (after Störmer).

The geographical position of the two aurorae which occurred respectively on (*a*) March 24, 1936, and (*b*) January 24, 1936, is shown in figs. 2 (*a*) and 2 (*b*). By measuring the photographs taken at the two stations the height and geographical position of several points on the auroral drapery were measured and are shown in fig. 2 (*b*) as the line marked 82. This drapery lay over the sea to the North of the Shetland Isles. Photographs taken nineteen seconds later showed that the drapery has shifted southwards to the position marked by line 83. In the case of the auroral arc (*b*) with ray structure (RA) photographs could only be taken at Trondheim as the sky was cloudy at Störmer's other auroral stations. Its position had therefore to be calculated from assumed heights (H) of 85, 95 and 105 km. Figure 2 (*a*) shows that they all lead to concordant positions of the arc partly over Norway and partly out at sea and illustrate the great heights of the arcs.

In fig. 3 are shown the horizontal projections of homogeneous arcs as derived from photographs taken at Oslo and Bossekop. The dotted lines are the circles of magnetic latitude, i.e. circles of equal angular distances θ from the pole of the uniform magnetization (not the north pole as indicated by a dip-needle). The interesting feature of this figure is the fact that the arcs very nearly coincide with these circles of magnetic latitude ; the small systematic deviation is about 10° according to Vegard.

§ 5. Low Aurorae and Auroral Sounds

Notwithstanding the fact that the Norwegian observers have given auroral heights exceeding 80 km, many observers claim to have seen aurorae below the clouds or near the ground. Again, whilst one of the characteristics of aurorae is the absolute silence during display, some observers have reported 'swishing' or 'rustling' sounds. Simpson (1905, 1933) discounts such reports and ascribes such impressions to fog or mist illuminated from above. In a review of observations of low aurora and auroral sound Chapman (1931, 1932) does not discard their possibility entirely ; indeed there is some evidence of sounds connected with strong aurorae heard by trained observers ; it is possible that they may be due to an indirect influence of the aurora though no physical explanation has been given.

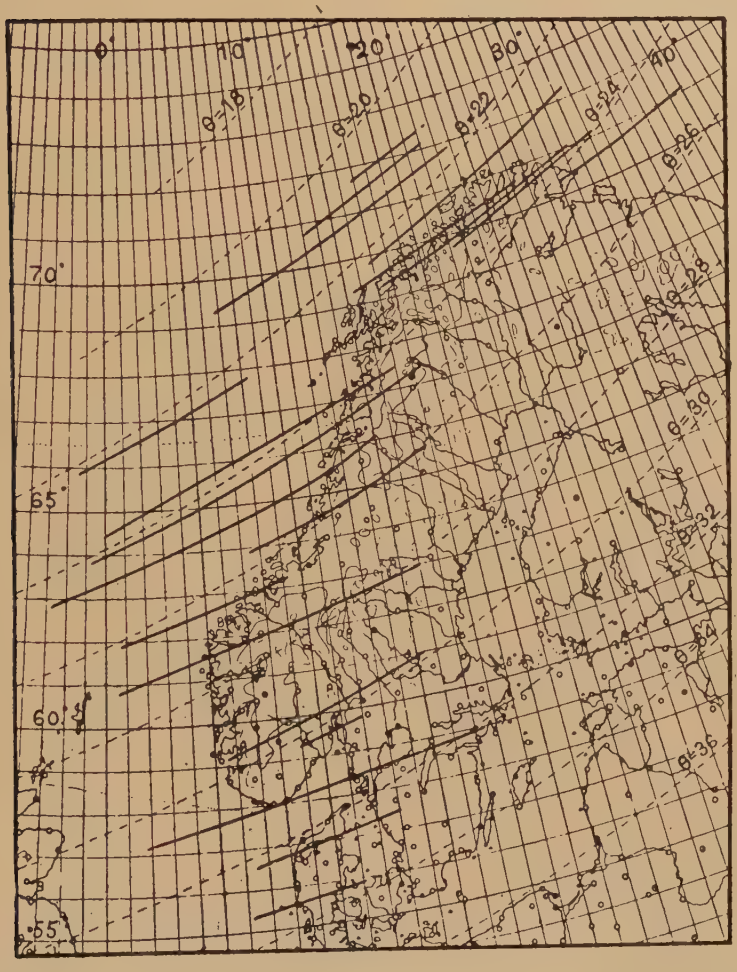
§ 6. Geographical Distribution of the Aurorae

The aurorae are most frequently seen near the arctic and antarctic regions, but they are occasionally seen over the greater part of the globe, especially during periods of magnetic disturbances. In non-polar latitudes, aurorae are seen on the poleward side of the zenith but further north they appear brighter and nearer to the zenith.

In 1881 Fritz computed from his catalogue of aurorae the average relative frequency of nights with aurorae, expressing it as M nights per year. These data were reduced to a common epoch to eliminate year to

year variations in frequency. Fritz drew a chart (fig. 4) showing lines of equal aurorae frequency M which he called *isochasms*. Vestine (1944) has revised and improved Fritz's diagram using subsequent data. The lines (beginning with $M=0{\cdot}1$, i.e. one aurora in about 10 years) are very nearly circular, with an ' auroral zone ' or line of maximum auroral frequency of mean angular radius of 23°, and having its centre at the pole of the geomagnetic axis. The dotted line close to this connects points at which aurorae appear equally often in the northern and southern sky.

Fig. 3



Showing the geographical position and direction (in plan) of a number of homogeneous auroral arcs observed by Störmer. Their direction agrees approximately with that of the circles (broken lines) of geomagnetic latitude (θ=angular distance from the north pole of the geomagnetic axis).

Observations of the southern lights are not so numerous as those for the northern lights, because of the paucity of antarctic stations. The southern ' aurora zone ' has been considered by Davies (1930) and also by White and Seddes (1939). It is shown in fig. 4 as of roughly circular shape

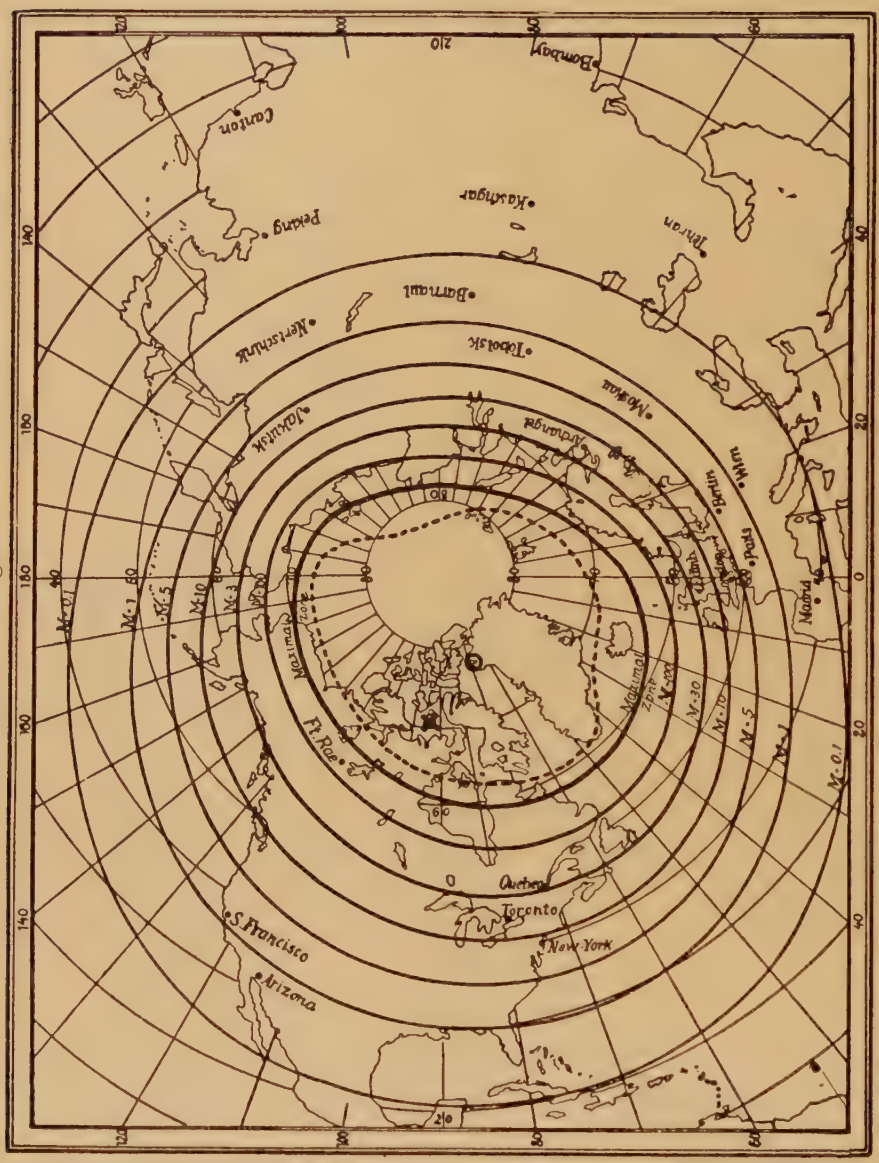


Fig. 4 The distribution of isochasms, or lines of equal auroral frequency, in the northern hemisphere, according to Fritz.

with the southern pole of the geomagnetic axis as centre. The radius of the zone can only be estimated with limited accuracy and seems to be about 18°.

§ 7. Diurnal and Annual Variation

In non-polar regions the aurora is most frequently seen in the early night hours with a maximum which occurs at different times at different stations, but usually one or two hours *before* local midnight (fig. 6).

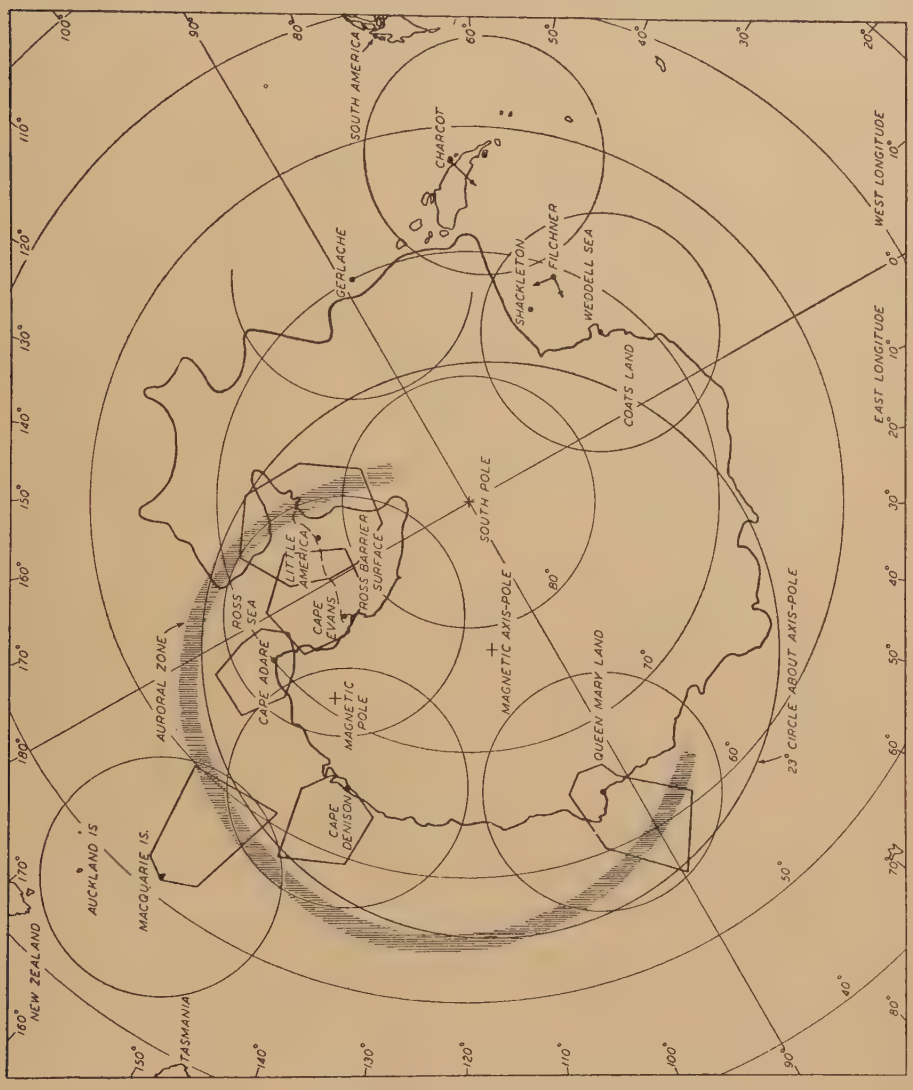


Fig. 5 Auroral zone in south polar regions.

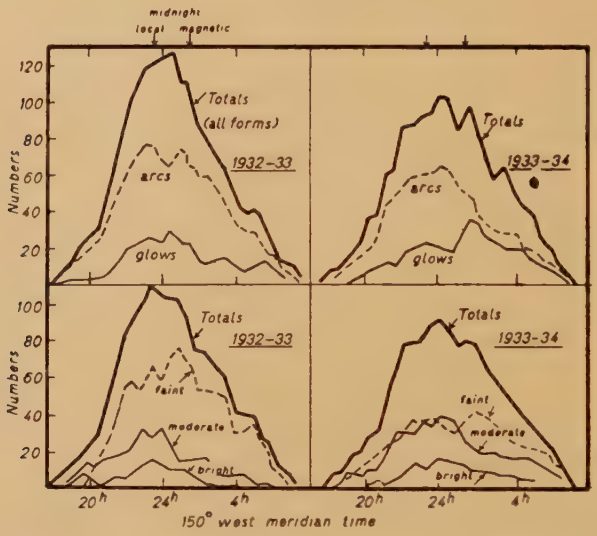
According to Harang (1950) this maximum seems to occur at 1·3 hr. before local magnetic midnight, i.e. the time when the plane through the station and the magnetic axis passes through the sun. Several stations record a second maximum, but the reality of this appears doubtful (Hulburt 1931).

An analysis of long and consistent auroral observations shows that the more brilliant forms occur before the times of maximum auroral frequency, faint and quiet glows being more frequent near morning.

The aurora also show a daily variation of position and direction and have a tendency to move southwards during a display (cf. § 5).

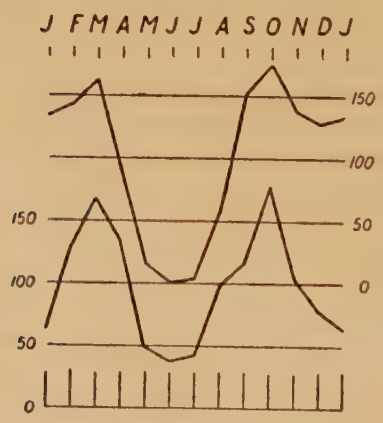
The curve of annual variation of aurorae shows two maxima at the equinoxes as in the corresponding curve of magnetic activity. This is shown in fig. 7 for Sweden. At polar stations this variation is less marked; a distinction between strong and weak displays should be made to get more homogeneous data.

Fig. 6



Diurnal variation of different forms and aurorae of different intensities at College, Alaska (after Fuller).

Fig. 7



The annual variation of auroral frequency, for Sweden (upper curve) and for the southern hemisphere (lower curve). The scale is adjusted so that the annual mean ordinate is 100 for each curve.

§ 8. Sunlit Aurora

Whilst the upper limit of arcs does not often exceed heights of 150 km rays extend to extreme heights of 400 km or more. On March 22 and 23, 1920, rays were observed as high as 800 km, their base being 400 km. On September 8, 1926, an aurora curtain, of violet-grey colour, stretching over the Shetland Isles, was seen from Oslo at a height of 300–500 km ; it later transformed into a diffuse mass extending to 1000 km.

Fig. 8

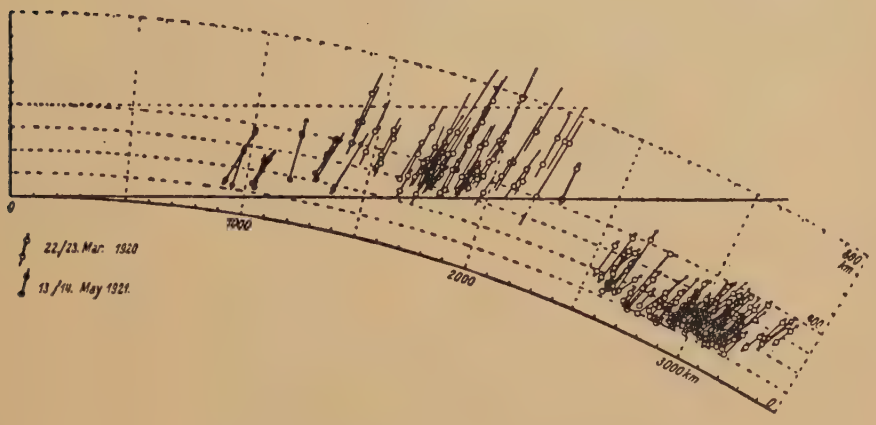
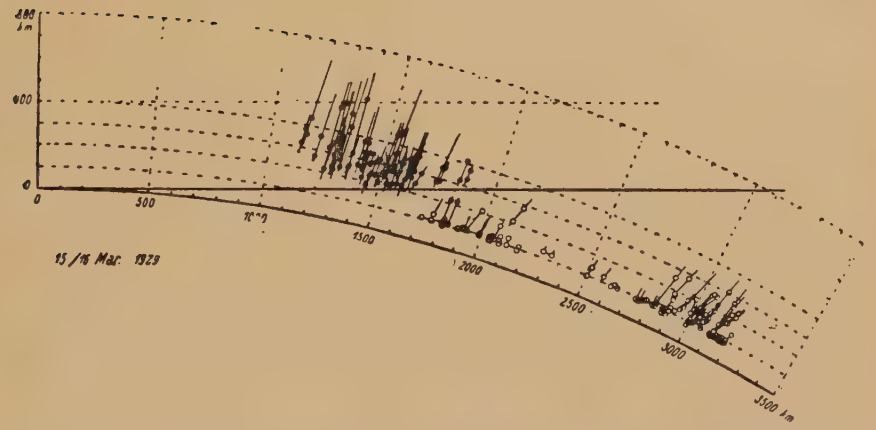


Fig. 9



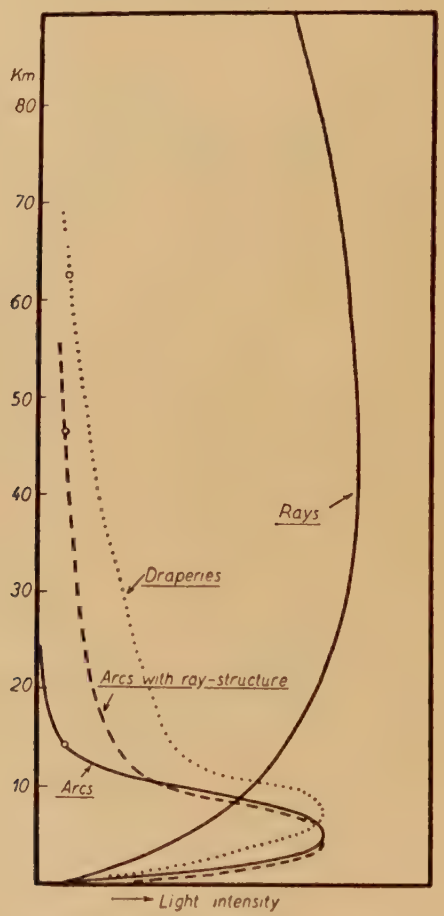
Diagrams illustrating the heights, and the positions relative to the earth's shadow-boundary, of auroral rays observed by Störmer : showing the exceptional heights attained by some sunlit aurorae.

Störmer (1929 a, b, 1930) found that these very high rays (above 400 km) are situated in the sunlit part of the atmosphere, over parts where the atmosphere is in the earth shadow up to 400 km. One characteristic

feature of the rays is that some consist of two luminous parts, one situated in the sunlit portion of the atmosphere, the other in the dark portion along the *continuation of the same line* (fig. 9). The ray is thus divided by a dark segment along the earth's shadow boundary.

As is to be expected for long arcs which lie partly in the sunlit and partly in the dark atmosphere, the lower borders of the arc is lower (100 km) in the portion in the atmosphere in shadow and higher (120–130 km) in the portion of the arc in the sunlit atmosphere.

Fig. 10



The mean variation of light intensity in vertical direction for different auroral forms. o indicates the upper limit of photographic impression (after Vegard).

The most outstanding difference between sunlit and ordinary types of aurorae is the height at which they occur, the latter rarely extending above 400 km as we have already said. The greyish-violet, and even blue colour on some occasions of sunlit aurorae is also a feature distinguishing them from ordinary aurorae.

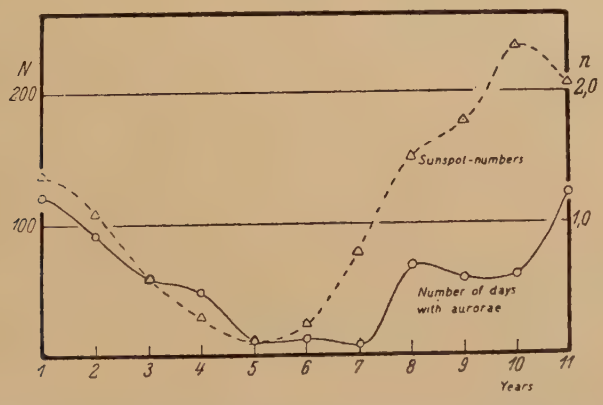
Nevertheless, on the night of January 25–26, 1938, during a very great geomagnetic storm, when the atmosphere appears to be blown up to unusually great heights, Störmer (1942) measured rays in the dark atmosphere extending up to 700 km.

§ 9. Variation of Intensity of the Aurora with Height

The intensity of the various forms of the aurora with height has been considered by Vegard (1921) and by Vegard and Krogness (1920). The results of the latter are shown in fig. 10 where the height is measured from the lower edge. The interesting feature of this diagram is that, except for aurorae with ray structure, the intensity rises rapidly to a maximum and then decreases more slowly, the luminosity thereafter becoming faint with increasing height.

The luminosity of the rays decreases only gradually with height ; their cross-section is sometimes very small amounting to not more than 300–400 metres in diameter.

Fig. 11



Sunspots (N) and auroral evenings (n) during the three sunspot cycles 1897–1907, 1908–18 and 1919–20. Place of observation : Denmark between 54° and 57° N. (after Egedal).

Connection with Solar Activity and Geomagnetism

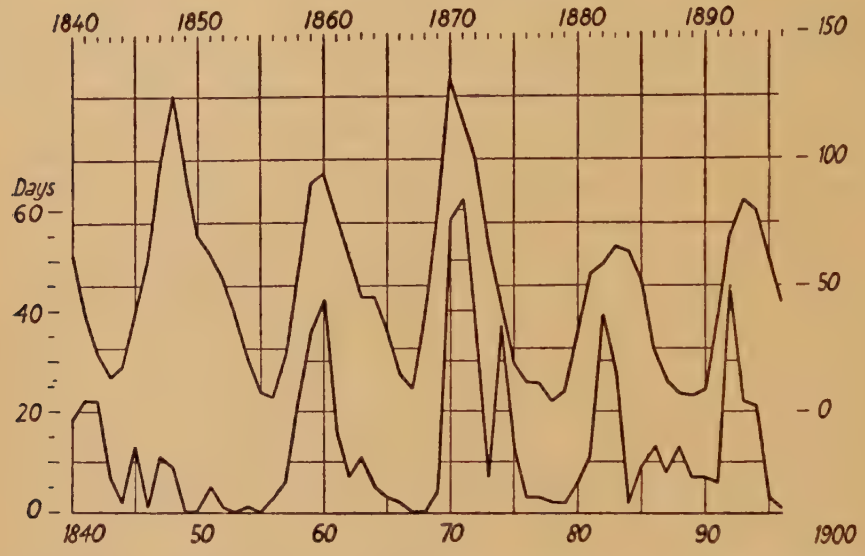
§ 10. The 27-day Recurrence Tendency

There is a definite tendency for aurorae to recur after a period of nearly 27 days, the period of the solar rotation. This was first noted by Fritz and also by Sverdrup (1927) during the expedition of the *Maud* ; Sverdrup found that this recurrence tendency was not confined to auroral displays of marked intensity. The connection between sunspots and aurorae was illustrated by Egedal who compared the auroral frequency with the sunspot numbers during the three solar cycles extending from 1897 to 1929. The results are exhibited in fig. 11 and suggest that there were big differences in the intensity of the aurorae in different 11-year periods.

§ 11. The Eleven-year Cycle

This connection is also brought out in fig. 12, after Boller (1898), which shows the 11-year cycle in the aurora australis : apart from the years 1840–60, there is a close connection with the curve giving the relative sunspot number. During the sunspot maxima about 1859 and 1870, some displays were seen nearly all over the globe ; one of the strongest displays occurred on February 4, 1872, when the aurora was seen as far south as Bombay (at 80° from the magnetic axis) and Aswan and reached the zenith at Constantinople and Athens.

Fig. 12



The number of days per year on which southern lights (aurora australis) were seen, in 1840–96 (lower curve) compared with the relative sunspot numbers (upper curve) (after Boller).

§ 12. Connection with Geomagnetism

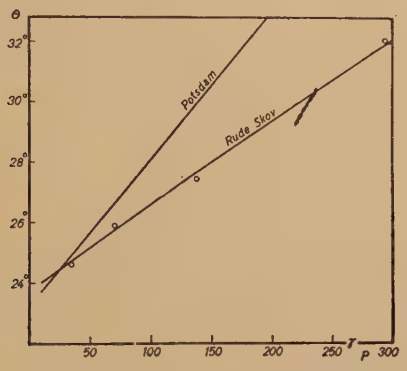
We have already mentioned the fact that the auroral rays lie along the magnetic lines of force of the earth's field. Further, the mean direction of the arcs is very nearly parallel to the magnetic parallels of latitude. Some connection between the secular variation of the earth's field and the direction of the arcs is also probable.

The close connection between aurorae and geomagnetic storms has been known from very early times and the fact was mentioned by Halley in 1716 ; but our knowledge of the auroral forms and the geomagnetic storm-time field is as yet incomplete. Strong auroral displays always accompany intense geomagnetic storms, and near the auroral zone the magnetic variations are especially difficult to analyse.

During an intense magnetic storm the auroral zones are widened considerably and, in the northern hemisphere, aurorae appear near the tropics (cf. end of § 11). During the storm of January 26, 1938, aurorae were visible over the greater part of Europe and yet in Tromso only a faint display was seen. Röstad (1935) has considered the increase in the angular distance θ of the most intense aurora from the geomagnetic axis with increasing magnetic perturbation at Rude Skar and Potsdam, and his results, shown in fig. 13, are of interest in this connection (cf. table 4, § 38).

Earth-currents are also closely connected with aurorae, as was to be expected from their own close relation with geomagnetic storms. During auroral displays Rooney (1934) reports amplitudes of earth current

Fig. 13



Widening of the auroral zone during strong earth-magnetic storms. θ is the maximum angular distance of aurorae from the magnetic axis point and P the magnitude of the perturbing vector at Rude Skov and Potsdam (after Røstad).

oscillations as high may be as 300–400 millivolt/cm. Near the auroral zone the earth currents in long telegraph cables are sufficient to cause extensive damage.

Lee (1930) at Lerwick has considered statistically the connection between auroral frequency and the daily magnetic character figure C. (This is defined as follows : all magnetic observatories classify Greenwich days with the figure 0, 1, 2 according as the day is relatively quiet, slightly disturbed or disturbed. The average of these figures for each day is defined to be the magnetic character figure.) For this purpose he derived *auroral character figures*, A, for each day ; the figure 0 being assigned when no aurora was observed during a favourable 3 hours period, the figure 1 when auroral forms occurred with no ray structure, the figure 2 for auroral forms with ray structures or also flaming aurorae. The results for 367 days of reliable observations are summarized in the table below (given by Chapman and Bartels, *Geomagnetism*, p. 471).

The fact that on two nights with magnetic character 2 no aurorae were observed is not exceptional, as on these occasions in one case the storm had subsided before dusk and on the other the aurora was seen near by. However, Vegard (1916) has observed cases when a ray aurora was accompanied by only a slight disturbance below it.

As the table shows, in general, the forms with ray structures are more likely to be accompanied by magnetically disturbed conditions than are quiet homogeneous forms.

Table 1

Auroral character A	Magnetic character C			
	0	1	2	Mean C
0	107	43	2	0·3
1	31	104	13	0·9
2	5	39	23	1·3
Mean A	0·3	1·0	1·6	

II. THE AURORAL SPECTRUM

§ 13.

The hypothesis that the aurora is due to reflected or refracted light had soon to be discarded since strong polarization of the auroral light as would have been expected was not found. The aurora is therefore self-luminous.

The spectrum of the aurora was first observed by Angström in 1867 who found a strong line at about $\lambda 5570$ in the yellow green portion. Much of the early work on the spectrum of the aurora was done by Vegard at Bossekop during 1912–13 (Vegard 1928). Later, measurements of the many lines in the visible portion of the spectrum were made by Lord Rayleigh (1928 a, b, 1930), Slipher (1919, 1929), Kaplan (1934, 1936), Dufay and Gauzit (1938), and Störmer (1937, 1938 a, b). More recently, Vegard and Kvifte (1945) have summarized the results for the main emission lines from high-latitude aurorae.

These consist of

(i) the green and red forbidden lines of atomic oxygen at $\lambda 5577$ and $\lambda 6300$ respectively,

(ii) the first positive band of molecular nitrogen in the red and infra red region,

(iii) the second positive band of molecular nitrogen in the ultra violet region,

(iv) the Vegard–Kaplan bands of molecular nitrogen in the blue-violet and ultra violet region,

(v) the first negative band due to N_2^+ in the blue violet and ultra violet region.

In addition, the forbidden ultra violet doublet of atomic nitrogen at $\lambda 3466$ has been identified by Bernard (1948), whilst Nicolet (1948) considers that certain band systems of O_2^+ and NO may also be emitted. Furthermore, Dufay, Gauzit and Tching-Mao-Lin (1941, 1942) have established the existence of a forbidden green doublet of atomic nitrogen at $\lambda 5199$.

There are also a number of unidentified lines at $\lambda 5206$ and $\lambda 5240$. The former of these may be due to a metastable state of N, but this is doubtful.

§ 14. The Green-auroral Line and other Oxygen Lines

The most important line in the auroral spectrum is the green auroral line at $\lambda 5577$: its origin was for a long time unknown. This line is also seen in the spectrum of the night sky, which indeed shows many of the lines observed in the auroral spectrum. The green line of the night sky was first measured by Slipher (1919, 1929) and L rd Rayleigh (1928 a, b, 1930) who called this light the ' non-polar aurora '. The wavelength was very accurately measured by Babcock (1923) with an ordinary camera placed behind a Fabry–Perot interferometer. He obtained the value $\lambda = 5577{\cdot}35 \pm 0{\cdot}005$ International Å.

McLennan and Shrum (1925, 1928) successfully reproduced the green line in the laboratory from atomic oxygen by introducing also neon or helium in the discharge tube. Later it was found that the intensity of the green line could be increased enormously when only a small quantity of oxygen was mixed with an inert gas, especially with argon.

Interference measurements by McLennan established the identity of the oxygen green line with the polar and non-polar aurora.

In the presence of a magnetic field the line was found to give a Zeeman pattern of a singlet line, and was therefore ascribed to the forbidden transition between the low metastable terms $2p'D_2$ and $2p'S_0$ in the arc spectrum of atomic oxygen. Since for forbidden transitions the transitional probability is very small (an atom or molecule concerned may remain excited for a long time before radiation—a few seconds instead of only 10^{-7} to 10^{-5} seconds as for allowed transitions), the rarified gases of the upper atmosphere, for which the collision-interval between atoms or molecules is long, favour these emissions.

After the identification of the green auroral line, the existence of the red oxygen triplet at $\lambda 6300$, 6364 and 6392 was found by Sommer (1930) in the spectrum of the aurora and by Paschen (1939) in the laboratory.

§ 15. The Nitrogen Bands

The first identified lines of the auroral spectrum were the nitrogen bands. The first negative band of N_2^+ is formed by the ionization of nitrogen by electronic impacts ; the head of the leading negative band $\lambda 3914$ requires, in all, 19·6 volts for excitation. The other bands are those with heads at $\lambda 4278$ and 4708 ; these together with the band with head at $\lambda 3914$ are prominent lines in the auroral spectrum, and the strongest of all nitrogen lines.

The spectrum of neutral nitrogen molecules appears in the auroral spectrum in a number of bands of low intensity belonging to the second positive band of N_2. They have heads at λ3997 and 4059 and are most numerous in the spectrum. The spectrum of oxygen and nitrogen lines are shown in plates V, VI.

The second positive band spectrum of N_2 lies in the ultra violet region but its intensity is low. The Vegard–Kaplan band lies in the blue-violet and ultra violet region but it also is very weak.

§ 16. Intensity Variations of the Auroral Spectrum

Vegard (1930) and his collaborators have made an extensive study of the variations of the relative intensities of the lines and bands with height and they distinguish between two types of variations, viz. *altitude effect* and *type effect.*

(*a*) *Altitude Effect*

Vegard (1932, 1933 a, b, 1937, 1938) found that for auroral streamers the intensity of the green line at λ5577 decreases very markedly relative to that of the negative bands of nitrogen, the relative intensity of these amongst themselves remaining sensibly constant along the whole length of the streamer. On one occasion, for the aurora of March 11, 1923, the intensity of the green line decreased by nearly 40% relative to the negative band in a distance of 50 km up the streamer. This effect of the weakening of the auroral green line relative to the positive bands has been found independently by Harang (1950).

Again, Vegard and Tonsberg (1938) have shown that there is a considerable increase in the intensity of the red doublet, as compared with the green line, with increasing height. In polar aurorae, this relative increase often amounts to as much as 50–300% from the lower border to the upper border as compared with the green band λ5577, the increase being greater for auroral forms of greater vertical extensions such as rays and draperies.

(*b*) *Type Effects*

The comparison of diffuse forms and those having more distinct features have been summarized by Vegard and others (1932, 1933 a, b, 1937) as follows :—

(i) The intensity of the green auroral line relative to the negative hydrogen bands is *smaller* for diffuse forms than for bands and draperies.

(ii) The intensities of the oxygen red lines and the first positive bands of nitrogen are *greater* for the diffuse forms than for bands and draperies.

(iii) Lines and bands in the blue region of the spectrum tend to be more prominent in diffuse aurorae.

However, as Harang (1950) points out, it is difficult to make a formal distinction between altitude and type effects, since different auroral forms may be at different heights and have different intensities. This is specially true in the case of aurorae which lie in both the dark and the sunlit atmosphere.

Spectra of *sunlit* aurorae show a strong decrease in the intensity of the green line λ5577 as compared with the first negative nitrogen band of N_2^+. Störmer (1937, 1938 a, b) and Vegard (1936) have also obtained spectra showing a strong enhancement of the red oxygen doublet at λ6300, 6363 as compared with the green auroral line. This is easily seen in plate VI showing the spectrum of a sunlit aurora taken by Störmer on September 15, 1938, with a yellow green auroral curtain whose lower border was at a height of 92 km.

§ 17. Hydrogen-line Emission in the Auroral Spectrum

The absence from the auroral spectrum of lines due to hydrogen and helium has long been taken as evidence of their absence from the atmosphere. Occasional appearance of the hydrogen Balmer lines has been noted by Vegard (1939, 1940) in auroral spectra showing well defined lines at λ6563 and 4861.

During the intense auroral activity on the nights of August 18–19 and 19–20, 1950, a spectrum of an auroral arc in the magnetic zenith was taken by Meinel (1950) on the night of August 19–20 which showed arc H_2 emission strongly asymmetric to the violet. The spectrograph was pointed parallel to the magnetic lines of force so that any incident auroral particles would be approaching the spectrograph. The asymmetrical H_2 profile showed a maximum displacement towards the violet of 60 A, corresponding to a velocity of 2800 km/sec. The H_2 line viewed perpendically to the magnetic lines of force, however, showed no asymmetry or displacement but showed a broadening of 6 A. This same broadening was also shown in the symmetrical H_2 line viewed parallel to the magnetic field. This effect had previously been observed by Gartlein (1951) whose observations were made with the spectrograph pointed normal to the magnetic lines of force.

Meinel's observations are of considerable importance in providing for the first time definite evidence that protons of probably solar origin are entering the upper atmosphere with velocities of 2500–3000 km/sec.

§ 18. Excitation of the Auroral Spectrum

The evidence furnished by Meinel's observations that swift particles enter the upper atmosphere confirmed the belief long held by many geophysicists that the aurora was excited by the direct or indirect effect of the impact of such particles in the high levels of the atmosphere.

The experience of electrical discharges suggests that the primary incoming particles will eject electrons in sufficient numbers and with moderate energy from the atmospheric atoms and molecules they encounter to give rise to the greater part of the luminosity, though the difficulty of obtaining the observed auroral emissions in the laboratory might, at first sight, make it seem surprising that the auroral green line and other forbidden lines appear so prominently in the auroral spectrum. But the conditions obtaining in the upper atmosphere differ greatly from those in a discharge tube where most atoms and molecules are deactivated by

collisions with the walls. There are no walls in the upper atmosphere and the low collision frequency between atoms and molecules at such high altitudes favour the emission of these spectral lines because they allow the metastable states to run out their spectral ' life '. Bates (1949) notes, for instance, that Götz (1947) found that the strength of the doublet at $\lambda 5199$ may continue for *hours* after the end of auroral display at a high altitude of 200 km or more. The non-appearance of this doublet in high latitude aurorae may then be explained by the greater collision frequency which de-excites the atoms before they have had time to radiate.

Several theories of the source of the auroral emission have been advanced by Vegard (1932, 1933 a, b, 1937), Chapman (1937), Ta You Wu (1943), Mitra (1945), *inter alia*. Most of these envisage some collision process in which a secondary electron, freed from the atmospheric molecules by the incoming particles, gives up its kinetic energy to cause excitation, which is then followed by emission. The following are probable processes :—

(i) First and second positive and Vegard–Kaplan bands

$$(1)\quad \begin{cases} N_2+e\rightarrow N_2'+e \\ N_2'\rightarrow N_2+h\nu. \end{cases}$$

(ii) First negative band of N_2^+

$$(2)\quad \begin{cases} N_2+e\rightarrow N_2^{+\prime}+2e \\ N_2^{+\prime}\rightarrow N_2+h\nu. \end{cases}$$

However, as shown by Bates, Massey and Pearse, other processes must also be operative. For instance, if the first negative band of N_2^+ is due to process (2), then each photon emitted may give rise to a positive ion–electron pair, and in equilibrium the number of recombination and ionization must be very great. Moreover, as many of these processes give rise to excited particles, the intensity of the ensuing emission must be considerable. A number of reactions are therefore possible ; for example (Bates (1949)) :

$$(3)\quad \begin{cases} N_2^++e\rightarrow N_2'+h\nu \text{ :—emission in continuum} \\ N_2'\rightarrow N_2+h\nu \text{ :—first and second positive and Vegard–Kaplan bands.} \end{cases}$$

Also,

$$(4)\quad \begin{cases} O+e\rightarrow O^-+h\nu \text{ :—emission in continuum} \\ O^-+N_2^+\rightarrow O'+N_2' \text{ :—ionic recombination} \\ O'\rightarrow O+h\nu \text{ :—green and red forbidden lines} \\ N_2'\rightarrow N_2+h\nu \text{ :—first and second positive Vegard–Kaplan bands,} \end{cases}$$

and finally

$$(5)\quad \begin{cases} O_2^++e\rightarrow O'+O'' \text{ :—dissociative electronic recombinations} \\ \left.\begin{matrix} O'\rightarrow O+h\nu \\ O''\rightarrow O+h\nu \end{matrix}\right\} \text{ :—green and forbidden lines.} \end{cases}$$

The relative occurrence of the processes (1)–(5) depends on various factors, such as the energy of the secondary electrons, the degree of ionization and constitution of the atmosphere. The great variety of colours emitted and their variations with height are thus not surprising.

Bates (1949) points out that the comparative brilliance of the sunlit aurorae is due to resonance scattering of the N_2^+ ions along the path of the incoming particles according to the reaction

$$(6) \quad \begin{cases} N_2^+ + h\nu \rightarrow N_2^{+\prime}, \\ N_2^{+\prime} \rightarrow N_2^+ + h\nu. \end{cases} \qquad \text{(First negative band system.)}$$

This may also provide an explanation of the 'divided aurora' in which part of a ray lies in the sunlit atmosphere and the remainder in the dark atmosphere, the two portions being divided by a dark segment. In this case, the luminosity of the lower portion is caused by ordinary collision processes (1)–(5), which, however, are insufficient to cause luminosity in the dark middle portion. The upper portion illuminated by the sun is due to resonance scattering. Bates also points out the remarkable fact that certain *allowed* lines of atomic oxygen are so faint, though Meinel (1948) has identified some strong allowed lines of O and N in the spectra of low latitude aurorae. One would expect such emissions to arise from the reaction

$$\begin{cases} O + e \rightarrow O' + e \\ O' \rightarrow O + h\nu. \end{cases}$$

§ 19. The Temperature of the Atmosphere at Auroral Levels

In his study of the green line of the light of the night sky, Babcock suggested the possibility of using the width of the spectral bands in the aurora due to the Doppler effect to determine the air temperature of those levels. Using the nitrogen bands in the spectrum, Vegard and Tonsberg found evidence that the air temperature between the levels of 110 and 150 km above the ground is −30°C to −40°C. Babcock had found that the green line of the night sky light was sharp and narrow which also points to the absence of Doppler broadening and so to very low temperatures.

III. THE NATURE OF SOLAR STREAMS

§ 20. Corpuscular or Ultraviolet Light Hypothesis ?

A successful theory of Aurorae must explain (*a*) the geographical distribution and polar incidence, (*b*) the height and form, (*c*) the diurnal variation, (*d*) the spectrum, (*e*) the close relationships between aurorae and geomagnetic and solar phenomena. The close relationship with solar phenomena must be taken as strong evidence of the solar origin of aurorae and geomagnetic disturbance, though the solar agent may be in the form of electromagnetic radiation (ultra violet light), or corpuscular in nature.

The evidence on the whole favours the corpuscular origin which ascribes the aurorae and magnetic disturbances to the terrestrial effects of stream of corpuscles emitted from the sun. The most conclusive evidence is undoubtedly Meinel's observation of the Doppler shift of the H_α line observed in the spectrum of an aurora viewed along the magnetic lines of force (see § 17). Moreover, the following objections can be directed against a light hypothesis of the solar agent.

(i) The flares of intense ultra violet light observed on the sun and accompanied by radio fade-outs, only cause a temporary augmentation of the solar diurnal variation of the earth's magnetic field which are quite unlike the geomagnetic disturbances in showing no differences as between the night and day hemisphere. This was first pointed out by McNish (1937).

(ii) It is difficult to account for the polar incidence of the aurorae if the cause is solar ultra violet light falling on the sunlit hemisphere only.

(iii) The 27-day recurrence tendency of magnetic storms, which is also noticeable in statistics of aurorae frequency data (§ 11) is unfavourable to a light hypothesis. Maunder (1904, 1905, 1916) showed that the recurrence tendency was most easily explained by supposing that the sun emits streams of corpuscles from limited disturbed areas of the sun's surface extending over a period of one or more solar rotations, and that magnetic storms and aurorae are produced when such a stream encounters the earth. If the emission lasts for more than one rotation, the stream will encounter the earth more than once and this accounts for the 27-day recurrence tendency.

This implies that the solar agent travels radially outwards as a limited beam. But it would be difficult to suppose that a burst of ultra violet light from solar flares can give rise to such a recurrence tendency, unless it can be laterally limited as in a searchlight beam.

(iv) The time-lag between the central passage across the sun of the disturbed solar area thought responsible for the terrestrial phenomena and the occurrence of geomagnetic disturbance is also unfavourable to a light theory. Newton (1943, 1944) has studied the distribution of time intervals derived from the more intense solar flares and geomagnetic storms within a period of 0 to 3 days after the occurrences of the flares, and finds that by far the greater number of storms occur in a period varying from about 22 hours for great storms to 34 hours for storms of smaller intensities. This agrees well with the time-lag found earlier by Greaves and Newton (1928, 1929) and Abetti (1929) between the central meridian passage of active sunspots and the commencement of a great magnetic storm. (Maurain (1927) found a rather larger interval (2·5 days).)

As Newton has shown, statistics indicate that the active regions of the sun responsible for magnetic storms and aurorae, such as flares, is limited roughly to the central half of the disc. This would account for the fact that periods of magnetic calm may follow the passage of a large group of large sunspots.

§ 21. The Nature of Solar Streams of Corpuscles

The fact that aurorae appear mainly in high latitude and that auroral rays follow the magnetic lines of force led Birkeland (1896) to suggest that aurorae are due to charged particles emitted from the sun deflected polewards by the earth's magnetic field. He supported his suggestion by experimenting with cathode rays which he projected towards a small magnetized sphere (or terrella) and showed that their distribution had many similarities with the form and distribution of the aurora, such as the occurrence of two 'auroral zones', and the arrival of particles on the 'night' side of the earth.

The corpuscular theory of aurorae appears to have been first suggested by Goldstein in 1881, but Birkeland's experiments gave the first real insight into the physical cause of the aurorae. His work was continued by Störmer, Vegard, Harang and others. Birkeland supposed that the particles responsible for the magnetic storms and aurorae to be electrons of very high velocity, nearly equal to the velocity of light. Störmer and Vegard considered the possibility that they might be slower electrons or positive ions.

Schuster (1911) criticized Birkeland's theory mainly on the ground that unless the velocity of the particles approached the speed of light, any stream giving rise to an appreciable magnetic field would be dispersed by the mutual electrostatic repulsion of its parts on the passage from the sun to the earth. To meet Schuster's criticism, Birkeland supposed that the speed of the electrons differed little from that of light. Dauvillier (1932) also proposed a theory of magnetic storms and aurorae based on fast electron streams which are deflected poleward by the earth's magnetic field and then away again, reaching a height of one earth's radius above the earth where they are supposed to produce high-speed secondary electrons by ionization of atmospheric molecules ; an aurora is then attributed to the downward motion of these secondary electrons. Meinel's recent evidence for the entry of high speed protons in the earth's atmosphere during auroral activity, however, is less favourable to theories of the aurora based on solar electron streams.

In 1919 Lindemann, now Lord Cherwell, urged criticisms similar to those of Schuster to a theory of magnetic storms based on charged corpuscular streams proposed by Chapman in 1918, but added to his criticism the suggestion that the essential features of Chapman's theory might be retained if the charged stream were replaced by an ionized stream electrostatically *neutral*. This would overcome the electrostatic objections to the hypothesis of solar streams composed of charges of one sign only. Because fast electrons would have penetr. ting powers which are comparable with those of cosmic rays, and so penetrate to lower levels than the known levels of 100–120 km to which auroral rays penetrate the earth's atmosphere, it seems likely that the only type of solar corpuscular theory which can be postulated to explain the aurorae arc are those which suppose the streams to be electrostatically neutral.

§ 22. The Emission of Solar Corpuscular Streams

Lindemann (1919) considered that such a stream of corpuscles could be emitted from the sun by the action of radiation pressure on atoms of the solar atmosphere and that these would then become ionized and draw electrons after them in approximately equal numbers. He estimated their speed of emission to be of the order of 800 km/sec, and showed that the amount of recombination to be expected during the passage of the stream from the sun to the earth would be small.

Milne (1926) subsequently showed that solar atoms such as H, He, Na, Fe, Mg, and ions, such as Ca^+, Ti^+, Sr^+, that produce strong absorption lines in the spectrum will be subject to an outward radiation pressure if they happen to be moving vertically upwards. For, on account of the Doppler effect, the increased velocity causes the atom or ion to absorb in a wave length which is steadily shortening (relative to a system fixed in the sun) until it becomes exposed to the full radiation pressure of the continuous spectrum. It is thus accelerated away from the sun against its gravitational field. Milne showed that after a distance of a few solar radii, the atoms or ions proceed into space with a uniform speed of the order of 1600 km/sec. This would correspond to a time of passage from the sun to the earth of about 26 hours, which agrees well with the time lag found by Newton between the occurrence of an intense solar flare on the sun and the commencement of a magnetic storm on the earth associated with the flare.

Milne's process indicates a continual escape of atoms from the whole of the sun's surface ; however, in order to explain the 27-day recurrence tendency of magnetic storms and aurorae, it seems essential that the streams should proceed from restricted areas of the sun's surface. Kahn (1949) has considered the possibility of Lα radiation pressure from a solar flare as the cause of the emission of the solar stream, and concludes that the radiation pressure produced by an intensive solar flare is unlikely to accelerate particles to a final speed of 1600 km/sec in sufficient numbers to produce a storm. Kiepenheuer (1952), however, has shown that Milne's acceleration mechanism applied to solar flares can produce enough Ca^+ ions to account for the energy of a moderate geomagnetic storm. (Alfvén has considered a process of emission based on the diamagnetic action of a solar magnetic field on a neutral ionized gas. This will be considered in § 31.)

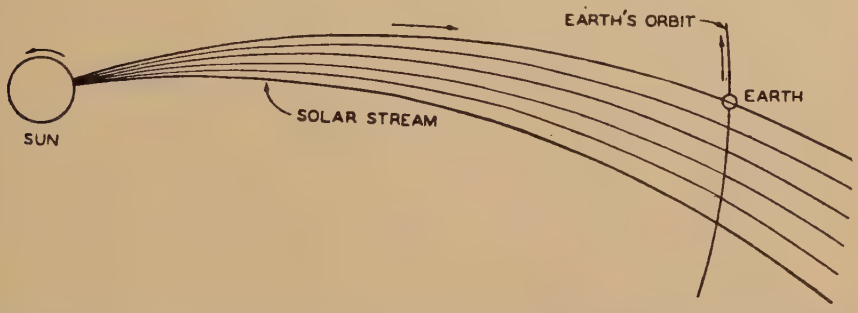
§ 23. The Detection of Solar Streams

Eclipse photographs show some evidence that the coronal streamers extend far out into space, the extension being often associated with disturbed areas on the sun. Eruptive prominences occasionally attain very high speeds. Allen (1938) has observed by measurement of the Doppler displacement, corpuscular emission directed towards the earth with a speed of 750 km/sec. But apart from theoretical studies, such as those mentioned above, we have little knowledge of the process of emission of a corpuscular stream.

Chapman (1929) suggested that evidence of the existence of solar streams could be obtained if these could be detected spectroscopically during magnetic storms. For speeds of emission of the order of 1000 km/sec the Doppler displacements of the principal absorption lines would be of the order of 15 to 20 A towards the violet side. The intensity of the absorption would depend on the total number of absorbing atoms per cm^2 column along the time of sight. The observations of such absorption lines are very difficult to make ; Richardson (1944) at Mount Wilson and Brück and Rutlantt (1946) at Cambridge have reported some success in the case of the H and K lines of ionized calcium. They indicated respectively speeds of the order of 700 km/sec and 1000 km/sec.

These results have been discussed by Chapman (1948) and Kahn (1949) who casts doubts on their reality, since the density of the stream required to produce an appreciable absorption line appears to be excessive in the light of current theories of magnetic storms and aurorae. Kahn considers that the intensity of the absorption lines may well lie below the observational limit. Furthermore, a stream consisting of protons and electrons would show no absorption lines and so could not be so detected.

Fig. 14



Sketch illustrating a solar stream of particles moving with a speed corresponding to a time of travel of 36 hours from the sun to the earth.

§ 24. The Geometry of Corpuscular Streams

The geometry of a stream of corpuscles emitted continuously from a small region of the sun's surface has been considered by Chapman (1929). Assuming an emission process such as Milne's, Chapman showed that the limiting velocity would be quickly acquired and that for limiting speeds of the order of 100 km/sec or more, the path of the particles would be very nearly rectilinear. However, the curve of the stream would be bent backwards, lagging behind the sun's rotation, as in the case of water issuing from a garden hose (fig. 14). Due to geometrical broadening the area of the cross-section will increase as the square of the distance, very nearly,* and at the distance of the earth's orbit it will be 46 000 times the area of emission. If there is no velocity spectrum along the

* Expansion due to thermal expansion and other causes is negligible.

stream the particle density will decrease inversely as the square of the distance and near the earth will be reduced to 1/46 000 times the density at emission. The composition of the solar stream is likely to be typical solar atmospheric gas, mostly hydrogen atoms (ionized to protons and electrons) with small amounts of other gas, such as ionized calcium. If the emission is from photospheric layers, the number density may be about 10^9 ions/c.c. and the temperature 6000°. If the emission is from lower levels, the number density will be less, and the temperature greater.

The value of the solid angle of a solar corpuscular stream at emission may be inferred if the annual variation of magnetic disturbances and aurorae frequency are attributed to the changing position of the earth with respect to the sun's equator. The sun's axis of rotation is inclined at an angle of 7·3° to the normal to the ecliptic so that the earth's heliographic latitudes vary between −7·3° and +7·3°. Considering a conical stream of semi-vertical angle α emitted radially from an active region on the sun's surface in heliographic latitude l, the stream will sweep over any distant object such as the earth whose heliographic latitude l' lies between $l \pm \alpha$. Most of the emitting areas on the sun are likely to lie on the same latitude as sunspots, which are most frequent between latitudes 10° and 15° (N. and S.) though they may appear in latitudes of ±30° or more ; and Newton (1943, 1944) places these emitting regions in an area of angular radius 45° centred at the centre of the disc. However, taking $l = 12\frac{1}{2}°$ as an estimate of the mean latitude of the emitting areas, it follows that the semivertical angle of the conicals streat at emission must not be less than 25° if $l' = 0$ (which occurs at the solstices) and not less than 11° if $l' = 7{\cdot}3°$ N. for areas in the northern hemisphere of the sun. This question has also been discussed by Gnevishev and Ol (1946).

§ 25. The Electrical State of Corpuscular Streams

Selective radiation pressure on atoms and ions far exceeds the general radiation pressure on electrons in the sun's atmosphere so that in the case of the emission of ions, the head of the emitted stream will be positively charged. However, electrons will be drawn from the solar atmosphere by electrostatic attraction tending to neutralize any resultant positive charge on it. Ferraro (1930) has shown that the differential velocity of the ions and electrons to be expected in this case is so small that there is little possibility that solar streams carry electric currents.

Moreover, any resultant charge would be quickly dispersed to the surface of the stream since an ionized gas is a very good conductor of electricity. It is well known that in a conductor the time taken by a volume distribution of charge to diffuse to the surface is of order $\tau = (4\pi\sigma)^{-1}$ where σ is the conductivity of the gas in e.s.u. In terms of the coefficient of diffusion, D, for a mixture of singly ionized positive atoms and electrons this is (Chapman and Ferraro 1929)

$$\sigma = \frac{e^2 N D}{kT}, \qquad \ldots\ldots\ldots \quad (1)$$

where e is the electronic charge in e.s.u., T the temperature of the gas, N the number density of the electrons and k is Boltzmann's constant ($1{\cdot}37\times10^{-16}$ c.g.s.). The product ND is very nearly constant, and for a gas at a temperature of 6000°C, $10^{-18}ND$ varies only from 316 to 6·5 as N varies from 10^{-1} to 10^{9}. Taking $ND=5\times10^{18}$, and $T=6000$°C, we find $\tau=3\times10^{-14}$ sec, so that the diffusion of any volume-charge in the stream to the surface is almost instantaneous, independently of the size of the stream.

To examine the dispersal of electric charges by electrostatic repulsion once they have diffused to the surface, we consider for simplicity the case of a uniform spherical cloud of ionized gas. Let Q be the total charge diffused and consider a surface in the layer of radius r concentric with the surface which contains a fraction θ of the total charge Q in the layer. The electric field at points of the surface is $\theta Q/r^2$ and the equation of motion of the charges in the surface is

$$m\ddot{r}=e\theta Q/r^2, \qquad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where m is the mass of the surface charge. Since initially $\theta Q\propto r^3$, it follows that $\ddot{r}\propto r$ initially, so that there will be no overtaking of charges in the surface by those immediately behind; the sphere of radius r will then contain always the same charge θQ and the solution of (2) is

$$v^2=v_0^2+(2e\theta A/m)(r_0^{-1}-r^{-1}), \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where v is the speed of the particle and v_0 its initial value. The initial thickness of the layer is neglected in comparison with the initial radius of the cloud r_0. By writing

$$r=r_0\cosh^2 u,$$

so that $u=0$ when $r=r_0$, (3) can be integrated to give

$$r_0(\sinh u\cosh u+u)=(2e\theta Q/mr_0)^{1/2}t, \quad . \quad . \quad . \quad . \quad (4)$$

where t is the time reckoned from the beginning of the expansion of the layer. Except at the beginning of the expansion, we may approximate (4) by using the fact that when r/r_0 is large, u is large, so that very nearly

$$r/r_0=(2e\theta Q/mr_0^3)^{1/2}t. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

If Q is expressed in terms of the fraction excess f of the number of charges of ion sign over the number of the other kind, then

$$Q=\tfrac{4}{3}\pi r_0^3fNe,$$

and

$$r/r_0=(\tfrac{8}{3}\pi f\theta N/m)^{1/2}et, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

a relation independent of r_0, the initial radius of the cloud.

Supposing, for example, that the excess charge on the cloud is positive, and that the ions are singly ionized calcium atoms, then after 10 minutes, 90% of the excess charge ($\theta=0{\cdot}01$) will be beyond a radius r such that $r/r_0=10^4(fN)^{1/2}$. If f is only 1/100 and N is 10^5, $r/r_0=3\times10^5$. This would represent a reduction in the density of charge less than 10^{-16} of

the original value and illustrates the rapidity with which the electrostatic repulsion disperses the charges in the stream and the great extent of their dilution.

It is also of interest to consider a spherical cloud initially uniform and consisting of charges of one sign only, as in the auroral theories of Birkeland and Störmer.

The results obtained above can be applied in this case if we replace Q by $\frac{4}{3}Ner_0^3$; eqn. (6) then becomes

$$r/r_0=(\tfrac{8}{3}\pi N/m)^{1/2}\,et.$$

If $N=10^5$ and the charges are electrons, then after 10 minutes r/r_0 will be about 10^{10}, indicating an even greater dispersion during the passage of the cloud from the sun to the earth.

The results show that the interior of an ionized stream of corpuscles must be electrically neutral to a high degree of approximation.

§ 26. The Penetration of Solar Corpuscles in to the Earth's Atmosphere

The velocity of emission of solar particles given by Milne's theoretical process of selective radiation pressure is of the order of 1600 km/sec, a value which accords well with the statistical estimates of the time of passage of the particles from the sun to the earth.

There is a difficulty, however, to which Milne drew attention, in explaining how particles with this speed can penetrate down to the observed auroral levels of 100 km in height or lower. If these heights mark the limit of penetration of incoming particles, they must have speeds considerably in excess of the velocity of 1600 km/sec derived by Milne, though we have as yet no definite information regarding the equivalent range R in air at S.T.P. Milne in 1926 estimated R for calcium ions with a speed of 1600 km/sec to be 0·15 cm, whereas, according to calculations by Chapman and Milne (1920) the equivalent amounts of air above 100 km and 80 km are about 0·4 and 5·6 cm, assuming that hydrogen is absent. Chapman (1929) considers that even these estimates should be increased by a factor of about 4 and 2 respectively to about 1·5 and 10 cm respectively, and the latter exceeds the range of the fastest α-particles known. Unless the estimates of the gas density in the upper atmosphere are lower than is generally supposed, we must suppose that auroral particles have speeds considerably greater than 1000 km/sec.

Estimates of R can be made using laboratory data collected by Das Gupta and Ghosh (1946); the results for electrons and protons are summarized in the following table 2 prepared by Bates (1949).

The table shows that protons penetrate to lower levels than electrons, and that to account for the penetration of protons to auroral levels their speeds must be about *ten times* the speed postulated by Milne and derived from the statistical time lag. This is in general accord with the speed of auroral particles, of the order of 3000 km/sec, inferred by Meinel (1950) from measurement of the Doppler displacement of the auroral H_α line.

On the other hand, electrons with speeds differing from the speed of light by only 10 metre/sec, postulated by Birkeland, would meet the difficulty that they would penetrate to levels far lower than the observed auroral levels. The possibility that some of the charges in a neutral ionized stream responsible for the aurora may be accelerated near the earth by electronic fields to speeds much higher than those which they acquire by emission from the sun is discussed in §§ 33 and 38.

The question of penetration of and ionization by auroral particles has recently been discussed by Sugiura, Tazima and Nagata (1952), who estimate the density of incident protons required to maintain the ionized state over the auroral zone to be 1 proton or less per c.c.

Table 2

Speed	Electrons Energy	Range R	Speed	Protons Energy	Range R
($\times 10^4$ km sec^{-1})	($\times 10^5$ ev)	(cm air STP)	($\times 10^4$ km/sec)	($\times 10^6$ ev)	(cm air STP)
1	0·0003	0·01	0·10	0·005	0·01
3	0·0026	0·04	0·25	0·033	0·05
6	0·011	0·2	0·5	0·13	0·2
12	0·047	3·4	1	0·52	0·8
18	0·13	18·0	2	2·1	7·5
24	0·34	83·0	3	4·7	30·0
27	0·66	220·0	5	13·0	190·0
29·7	3·1	1300·0	10	53·0	3000·0

IV. THEORIES OF THE AURORA

(*a*) Corpuscular Theories of One Sign

§ 27.

In §§ 20 and 21 we have seen that, on the whole, the evidence favours corpuscular theories of the aurora rather than light theories. The earliest corpuscular theories, those of Birkeland, Störmer and Vegard, supposed that the streams consisted of charges of one sign; they showed many interesting analogies with auroral phenomena and gave a first real insight into their physical cause. Because of the neglect of the electrostatic forces, however, it seems unlikely that they will prove correct though some features of it may find a place in the ultimate theory of aurorae. It is gratifying that Störmer's theory has found a permanent place in the theory of cosmic rays.

Dauvillier's theory is also unsatisfactory, because, *inter alia*, of the discordance of (i) the level of penetration with the observed height of the aurorae, (ii) the time lag between the occurrence of solar flares and magnetic storms.

§ 28. The Birkeland–Störmer Auroral Theory

This theory ascribes the auroral phenomena to the action of the geomagnetic field on a solar corpuscular stream of charges of one sign. Birkeland attempted to infer the solution by projecting a beam of electrons towards a small magnetized sphere or terrella, and showed that the motion of the electrons was guided by the magnetic field of the terrella towards the polar regions, leaving a toroidal space free of electrons round the equatorial regions of the terrella.

Störmer (1904, 1911–12) developed the theory mathematically but restricting his investigation to the motion of a single charge in the geomagnetic field.* He has shown that such a particle emitted from the sun and travelling into the earth's magnetic field can only reach the earth along two narrow zones centred round the two poles of the earth's axis. These Störmer identified as the auroral zones. The angular radius of the zones depends on the value of e/mv for the charges considered, where e, v and m are respectively the charge, velocity and mass of the particle.

Because the angular radius of the zones for the various kinds of charged particles considered by him did not agree with the observed value of about 23°, Störmer sought to overcome the difficulty by postulating a great circular ring of electric charge revolving round the earth in its equatorial. The magnetic field of this ring current is supposed to enlarge the angular radius of the zone; at the same time it was thought responsible also for the great increase in the radius of the auroral zones during magnetic storms (see § 29).

Using the results of a great many computations of trajectories from the sun, Störmer also showed that a conical stream of corpuscles of small solid angle from the Sun would on arrival at the earth spread out along a band lying along the auroral zone, the thickness of the band being very small compared with its length; Störmer pointed to this as a possible explanation of the formation of auroral curtains.

§ 29. Störmer's Equatorial Ring-Current

Störmer did not express any views concerning the nature of the solar particles. But a ring analogous to the equatorial ring current postulated in the previous section had been observed by Birkeland in his 'terrella' experiment outside the forbidden toroidal regions. Störmer showed by a large number of detailed calculation that if a bundle of particles is projected from the sun in the earth's magnetic equatorial plane, many particles would flow partly round the earth, the negative particles flowing eastward and positive particles westward. Störmer imagined that these would form a ring of particles round the earth carrying a westward positive current. The mean radius of the ring current would be large

* Several excellent accounts of this part of the theory exist in the literature of the subject and, as this theory has now largely a historical interest, no detailed account of the mathematical theory will be given here.

comparable with the radius of the earth (cf. table 2) so that the effect of gravity can be neglected. The mean radius of the ring, r, can be determined from the equation of normal acceleration of a particle in the central filament of the ring, namely

$$mv^2/r = evH = evM/r^3,$$

where M is the magnetic moment of the earth. Hence

$$r^2 = Me/mv = l^2,$$

where l is called Störmer's constant. The values of l for various particles are given in table 3.

Table 3. Values of Störmer's Constant l in cm for Various Particles

$v=$	10^7	10^8	10^9	10^{10} cm/sec
Electron	$1{\cdot}2\times10^{13}$	$3{\cdot}9\times10^{12}$	$1{\cdot}2\times10^{12}$	$3{\cdot}9\times10^{11}$
H^+	$2{\cdot}9\times10^{11}$	$9{\cdot}0\times10^{10}$	$2{\cdot}9\times10^{10}$	$9{\cdot}0\times10^{9}$
Ca^+	$4{\cdot}5\times10^{10}$	$1{\cdot}4\times10^{10}$	$4{\cdot}5\times10^{9}$	$1{\cdot}4\times10^{9}$

The westward current in the ring produces a world wide, uniform decrease in the magnetic field of the earth at its surface, such as is observed during the main phase of a geomagnetic storm. The effect of the decrease of the magnetic field near the earth is (as was shown by Störmer (1904, 1911–12)) to increase the angular radius of the auroral zone. It is thus natural to suppose that the observed increase in the radius of the auroral zones during magnetic storms is due to the presence of a ring-current. Störmer went further and suggested that even at ordinary times the ring-current exists and enlarges the radius of the auroral zones.*

It is unlikely, however, that such a ring-current could be composed of charges of one sign only. The chief argument against it is that it could not hold together against the mutual repulsion of its parts, as has been indicated in § 25 (Schuster's criticism). Another unsatisfactory feature is that the ring would more than nullify the earth's magnetic field in the region where it is set up, thus reversing the forces which are supposed to set it up.

These difficulties, moreover, are *not* removed if, as has been suggested, we suppose the ring to be electrostatically neutralized by electrons drawn from outer space; for in this case there would be the added difficulty of a tendency for the current to be disrupted through constriction by 'pinch' effect resulting from the magnetic field of the current in the ring.

*This effect has been beautifully demonstrated experimentally by Brüche (1931).

These criticisms apply to the whole of Störmer's auroral theory, based on the fundamental hypothesis that the motion of the charged particle is governed solely by the earth's field, and independent of any interaction (electrostatic or otherwise) between the particles of the stream.

(*b*) Theories based on Neutral Ionized Corpuscular Streams : Alfvén's Theory

§ 30.

Auroral theories based on neutral corpuscular streams have been proposed by Alfvén (1939, 1940) and Martyn (1951) respectively. In Alfvén's theory of magnetic storms and aurorae the electrons in the stream have large energies—of the order of 10^8 e-volts at emission—but the ions need not have such large energies. The motion of any individual particle is supposed to be hardly affected by the magnetic field due to the motion of other particles.

A distinctive feature of the theory is the important part played by the solar magnetic field, both with regard to emission of the stream and the effects on arrival in the earth's field.

Martyn's theory of the aurora is based on the Chapman–Ferraro (1931, 1932, 1933, 1940) theory of magnetic storm and in particular, on the existence of a ring-current flowing in the earth's magnetic equatorial plane.

A common feature of both theories is that the aurora is attributed to the penetration in the earth's atmosphere of particles accelerated from the surface of the neutral stream in the neighbourhood of the earth by local electrostatic fields. Both differ from Störmer's theory in that the ring-current to which the equatorial disturbance of a geomagnetic storm is ascribed flows at distances of only a few earth's radii from the earth instead of several hundred (§ 29).

§ 31. The Motion of a Charged Particle in Crossed Electric and Magnetic Fields

In Alfvén's theory the motion of the charges is supposed to be entirely governed by the magnetic field of the sun and earth. In discussing their motions we follow Cowling (1942) rather than Alfvén.

Consider a particle of charge e (e.m.u.) moving in an inhomogeneous magnetic field $\mathbf{H}$. Take the origin near the particle at Ox in the direction of $\mathbf{H}$ at the origin. Near the origin the resolutes of H are approximately

$$\alpha_1 x+\beta_1 y+\gamma_1 z, \quad \alpha_2 x+\beta_2 y+\gamma_2 z, \quad \alpha_3 x+\beta_3 y+\gamma_3 z+H, \quad . \quad . \quad (1)$$

where $\beta_1=(\partial H_x/\partial y)$, etc. evaluated at the origin. Since div $\mathbf{H}=0$, we have

$$\alpha_1+\beta_2+\gamma_3=0. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Further, neglecting the effect of the currents, as in Alfvén's theory, **H** is derivable from a potential, so that

$$\alpha_2=\beta_2, \quad \alpha_3=\gamma_1, \quad \beta_3=\gamma_2. \quad \ldots\ldots (3)$$

The equations of motion of the charge are

$$m\dot{x}=e(\dot{y}H_2-\dot{z}H_z), \text{ etc.} \quad \ldots\ldots (4)$$

A first approximation is obtained by neglecting the variations in **H**, so that, by (1),

$$H_x=0=H_y, \quad H_z=H,$$

and by a suitable choice of the axes Ox, Oy, the solution is

$$x=(v_T/\omega)\sin\omega t, \quad y=(v_T/\omega)\cos\omega t, \quad z=v_H t, \quad \ldots (5)$$

where t is the time and $\omega=eH/m$; here v_T and v_H denote the resolutes of the velocity transverse and parallel to the magnetic field. This approximation is used to express the resolutes of **H** given by (1) as a function of t. Averaging by neglecting terms like $\sin\omega t$, $t\sin\omega t$, $t^2\sin\omega t$, we find that the mean velocity of the particle can be divided into two parts : the first part, called the inhomogeneity drift, u_I, has resolutes

$$-\frac{m}{eH^2}(\tfrac{1}{2}v_T^2+v_H^2)\frac{\partial H}{\partial y}, \quad \frac{m}{eH^2}(\tfrac{1}{2}v_T^2+v_H^2)\frac{\partial H}{\partial x}, \quad 0, \quad \ldots (6)$$

and is perpendicular to **H** and to the gradient of H. The second part of the mean velocity has resolutes

$$e\gamma_1 v_H^2 t/m\omega, \quad e\gamma_2 v_H^2 t/m\omega, \quad v_H-\tfrac{1}{2}e\gamma_3 v_T^2 t/m\omega, \quad \ldots (7)$$

and is very nearly parallel to **H**, which at the point $(0, 0, v_H t)$ has resolutes $\gamma_1 v_H t$, $\gamma_2 v_H t$, v_H ; and its magnitude gives the value of v_H after time t. Since this may also be written $v_H+\dot{v}_H t$, we have that the particle has a mean acceleration along the line of force about which it is spiralling of magnitude,

$$\dot{v}_H=-\frac{1}{2}\frac{ev_T^2 t}{2m\omega}\gamma_3=-\frac{v_T^2}{2H}\frac{\partial H}{\partial z}=-\frac{v_T^2}{2H}\frac{dH}{ds}, \quad \ldots\ldots (8)$$

where ds is an element of the line of force. This acceleration is towards a point of minimum H on the line of force.

Suppose that an electric field **E** e.m.u. also acts on the particle. Then the particle has an additional cross field drift u_E near the origin, perpendicular to both the electric and magnetic fields, of resolutes

$$E_y/H, \quad -E_x/H, \quad 0 ; \quad \ldots\ldots (9)$$

the acceleration $\dot{v}_H$ is altered to

$$\dot{v}_H=-\frac{v_T^2}{2H}\frac{dH}{ds}+\frac{eE_z}{m}=-\frac{v_T^2}{2H}\frac{dH}{ds}-\frac{e}{m}\frac{dV}{ds},$$

where V is the electric potential. But

$$\dot{v}_H=\frac{d}{ds}(\tfrac{1}{2}v_H^2)=\frac{d}{ds}\left(-\tfrac{1}{2}v_T^2-\frac{e}{m}V\right)$$

by the energy equation. Hence

$$\frac{d}{ds}v_T^2 = \frac{v_T^2}{H}\frac{dH}{ds}$$

and integrating we have $$v_T^2 = CH, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

where C is a constant along a magnetic line of force. The relativistic correction need not be considered since it is not needed in Alfvén's theory. It will be convenient to summarize here the four parts into which the motion of a particle has been divided. They are

(i) a circular motion, represented by (5), transverse to the magnetic field,

(ii) an oscillation along a line of force (Störmer oscillation) due to the fact that a charge is accelerated towards a point of minimum H,

(iii) an inhomogeneity drift, u_I, due to radial inhomogeneity of H, represented by (6),

(iv) a crossed-drift, u_E, represented by (9), due to the combined action of the electric and magnetic field and perpendicular to both.

§ 32. Alfvén's Emission Theory of Solar Streams

The process of emission considered by Alfvén differs from those discussed in § 22, in that the outward motion of the particles is ascribed to electromagnetic forces and not to radiation pressure. Alfvén assumed that

(i) the sun is surrounded by a magnetic field which may be represented by the field of a dipole of moment 10^{34} e.m.u. corresponding to a value of 50 gauss deduced by Hale for the solar magnetic field at the poles,*

(ii) electrons with energies of the order of 10^8 electron volts are emitted from limited areas of the sun's surface accompanied by an equal number of positive ions, so as to render the mass of ionized gas electrically neutral.

The energies of the positive ions need not be as high as that of the electrons. For these energies the radius of the circular motion of the electrons is small (10^4 to 10^5 cm) compared with the radius of the sun, so that we can apply the approximate analysis given in § 31. Hence the charges will acquire ' inhomogeneity drifts ', u_I, eastward for the electrons, westward for the ions, due to the inhomogeneous solar magnetic field. This will give rise to surface charges on the flanks of the stream, whose magnitude will be limited by the rate at which the charge in these layers is driven off by electrostatic repulsion. An eastward electric field E is therefore set up in the stream (fig. 15).† Alfvén has made a rough estimate of the rate of escape of the charges and hence of E and of the outward drift.

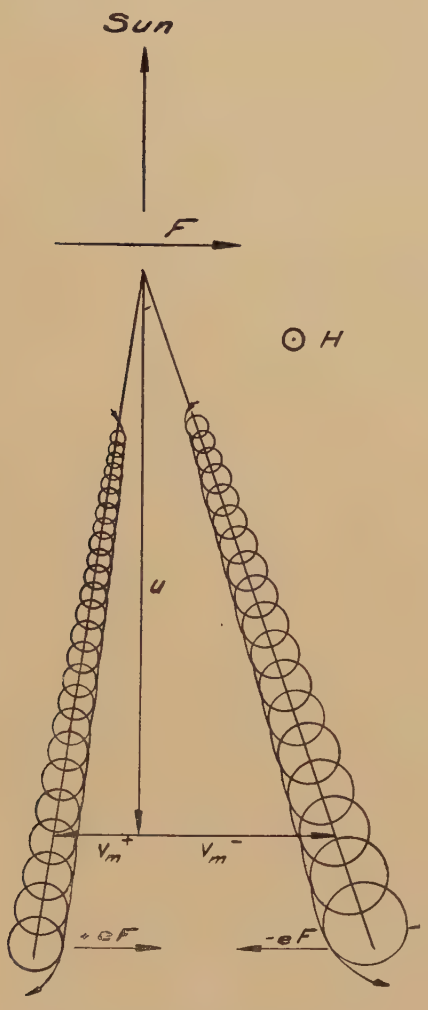
---

* There appears to be conflicting observational evidence for a permanent magnetic field of the sun of this order of magnitude. The most recent attempts to detect a general solar magnetic field show that if this exists, it cannot exceed a few gauss.

† This is denoted by F in figs. 15–19.

In conjunction with the solar magnetic field, the electric field E produces a ' cross-drift ', u_E, which is directed radially outwards and so carries the stream away from the sun.

Fig. 15



Because the particles oscillate along the lines of force about the sun's equatorial plane, whatever the latitude of the active areas responsible for the emission, the ionized stream will soon become symmetrical with respect to the equator and elongated in the north–south direction as compared with the east–west direction.

This implies that if such a stream causes a geomagnetic storm, then whatever the position of the source on the sun at emission, one would not expect occurrence of magnetic storms to be dependent on their

position. Bartels (1932) has, in fact, shown that there is an absence of correlation between the frequency of storms and the heliographic latitudes of the earth and the emitting area of the sun.

During the passage of the stream from the sun to the earth, the inhomogeneity drift carries the particles in a direction opposite to that of the electric force acting on them, so that they lose energy as they move into the weaker parts of the solar magnetic field, and in accordance with (10).

§ 33. Alfvén's Theory of Aurorae

The earth's magnetic field becomes comparable with the solar magnetic at a distance from the earth of the order of 3×10^{10} cm. If the duration of a magnetic storm of 1 to 2 days is taken to be equal to the time taken by the stream to sweep across the earth, the angular breadth of the stream (as viewed from the sun) is 15°–30°. At the distance of the earth's orbit, this corresponds to 4×10^{12} cm, so that the stream presents at the earth a large surface. Thus, only a small volume of the stream will be subject to the influence of the earth's field and according to Alfvén's this will not disturb the constant electric field in the stream produced by the inhomogeneity drift in the *solar* magnetic field. Once the stream enters the earth's magnetic field, the gradient of the combined magnetic fields of the sun and the earth will change sign (at a distance of the order of 6×10^{10} cm, if the solar magnetic moment is taken as 10^{34} ergs). Thereafter the electrons and ions will drift in directions opposite to those they have at the sun end. The positive charges now drift in the direction of the electric field E and hence the energies of the electrons and ions increase—the reverse process of the energy decrease during emission from the sun which takes place according to (10).

The problem is then to determine the motion of the particles in the stream under the influence of the constant solar electric field E and the earth's dipole field.

The increase in the potential energies of the electrons as they enter the earth's field means that the kinetic energy is increased ; this increase affects mainly the transverse component of velocity which is supposed large compared with the velocity parallel to the lines of force. It implies that the oscillations along the lines of force become small and that the stream tends to become flattened in the earth's equatorial plane.

The energy of the particle eV is then approximately given by*

$$eV/H=\text{const}=\mu. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

---

* This may be seen as follows : if the kinetic energy of a particle is denoted by T, and the resolutes of the drift velocity of the charge by u and v, along a stream line we have by energy consideration

$$\text{(i)} \qquad \frac{dT}{dt}=e(uE_x+vE_y)=-\frac{T}{H^2}\left(E_x\frac{\partial H}{\partial y}-E_y\frac{\partial H}{\partial x}\right),$$

using (6) and the fact that the cross-drift, u_E, is perpendicular to the electric field. Again, neglecting the magnetic field of the motion of the charges, and

The quantity μ has the dimensions of a magnetic moment—the *orbital magnetic moment* since it is due to the circular motion of the electrons about a magnetic line of force. This is supposed to be large compared with cross drift, u_E. We can then determine the trajectories of the electrons from eqn. (11).

Take the x-axis parallel to the electric field and assume their motion of the electrons is parallel to the xy-plane. The combined magnetic field of the earth and sun, parallel to the z-axis, is given by

$$H=H_0+\frac{a}{(x^2+y^2)^{3/2}}, \quad \ldots\ldots\ldots \quad (12)$$

where H_0 is the nearly homogeneous magnetic field of the sun near the earth and a is the magnetic moment of the earth. The energy of an electron at (x, y) is given

$$eV=eV_0+eE(x'_0-x), \quad \ldots\ldots \quad (13)$$

where x'_0 is the symptotic value of x at infinity. Combining (11) and (13), we have

$$\mu=\frac{eV_0+eE(x'_0-x)}{H_0+a(x^2+y^2)^{-3/2}},$$

or, say,

$$x_0-x=L^4r^{-3}, \quad \ldots\ldots\ldots \quad (14)$$

where

$$\left.\begin{aligned} r&=(x^2+y^2)^{-1/2} \\ L&=(\mu a)^{1/4}(eE)^{-1/4} \end{aligned}\right\}. \quad \ldots\ldots \quad (15)$$

Equation (14) gives the trajectories of the electrons and these are shown in fig. 16. It will be seen that the electrons stream past the earth on the morning side if $x_0<x_D$, but on the evening side if $x_0>x_D$, where the trajectory through $x=x_D$ is a dividing line. Consequently the electrons stream past the earth without penetrating an unsymmetric region (analogous to the ' forbidden regions ' in Störmer's theory). The positive ions in the stream describe similar trajectories except that the curves are mirror images of those in fig. 15 in the sun–earth line, but differing in scale according to the value of L for the positive ions.

Alfvén assumes that the temperature of the positive ions stream is considerably lower than that of the electron stream (§ 32), though (Alfvén 1950) he admits that this assumption is open to criticism. The value of L for the ions is thus much smaller than for the electrons, so that

---

the inhomogeneity drift compared with the cross-drift along a stream line, as is justified, we have

(ii) $$\frac{dH}{dt}=u\frac{\partial H}{\partial x}+v\frac{\partial H}{\partial y}=-\frac{1}{H}\left(E_x\frac{\partial H}{\partial y}-E_y\frac{\partial H}{\partial x}\right).$$

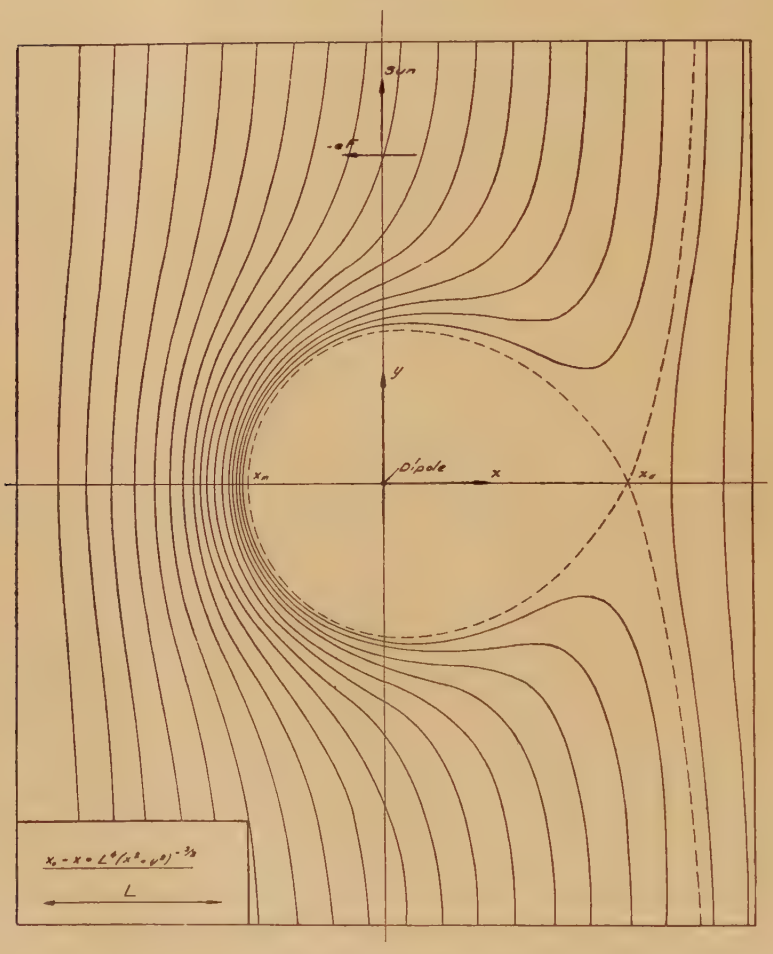
Comparing (i) and (ii), we have

$$T\frac{dH}{dt}=H\frac{dT}{dt},$$

and integrating $T=\text{const}\times H$ along a stream line. Expressing T in electron volts V, we have $T=eV$, whence $eV/H=\text{const}$ ($=\mu$ say).

the ' forbidden region ' for the positive ions lies entirely within the ' forbidden region ' for the electrons. This means that outside the ' forbidden region ' for the electrons the ions move in nearly straight paths parallel to the y-axis (see fig. 17). Alfvén has shown* that the density of either the electronic or ionic stream is inversely proportional to H, so that if the stream is electrically neutral at infinity it remains so

Fig. 16



Motion of the electrons.

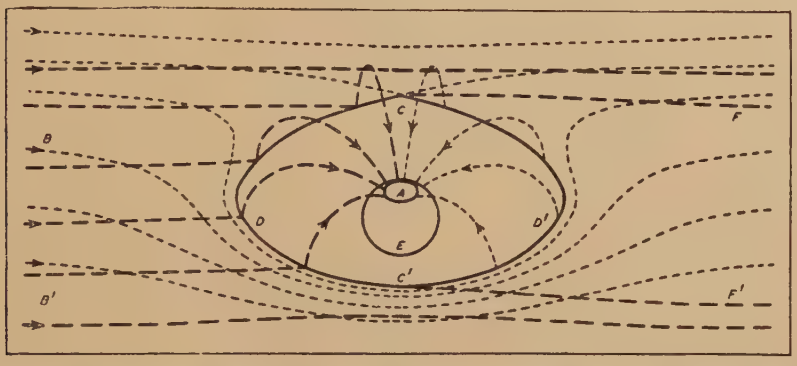
* The writer has verified this result independently and in a more general manner than Alfvén's; he takes this opportunity of withdrawing a former criticism of the theory to the effect that the stream would be unable to penetrate into the earth's field without violating the equation of continuity. This criticism would only apply if the dimensions of the stream are smaller than those of the region of space in which the gradient of the magnetic field of the earth exceeds that of the sun, as Alfvén pointed out in answer to this criticism.

throughout the region of space common to both ions and electrons. This means that the positive ions penetrate the ' forbidden region ' for the electrons on the morning side on arrival at its boundary and so create a positive space charge near the boundary CDC′ (fig. 17).

Because of the flattening of the stream into the earth's equatorial plane this is nearly a line charge. On the night side of the border the electrons are uncompensated by positive ions and a negative space charge is produced.

Alfvén then supposes that these space charges neutralize each other by a discharge along the magnetic lines of force over the polarcaps (fig. 18). The charges thus repelled from the boundary of the ' forbidden region ' are supposed to cause aurorae and the polar magnetic disturbances. On the night side of the earth the auroral particles are thus electrons.

Fig. 17



Perspective drawing of paths of ions and electrons near the earth (after Cowling); for simplicity, only the paths of charges reaching the north auroral zone are shown.

Paths of ions = — — —; paths of electrons = - - - - -; E = earth; curve A = north auroral zone; BB′ = ions and electrons approaching earth in equatorial plane; CDC′D′ = boundary of forbidden region; CF and C′F′ are boundaries of shadow of earth where no ions are present.

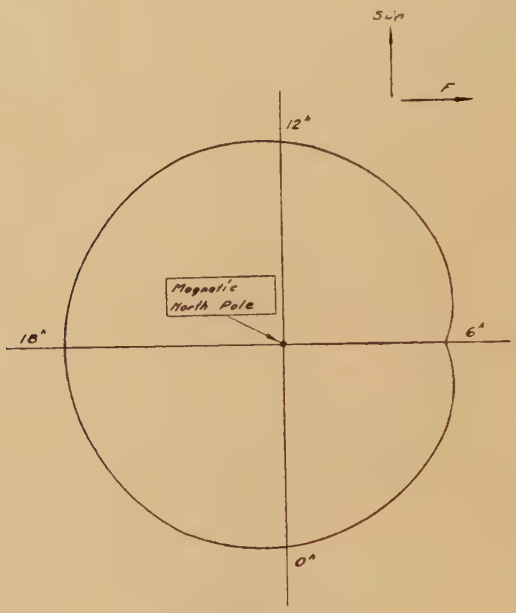
Projection of this boundary along the lines of force over the surface of the earth produces an ' auroral zone ' shown in fig. 18. The curve is nearest to the pole at 6^h and furtherest at 18^h so that if the observer has the zone at the zenith at 0^h and 12^h, then its situation changes from 1·4° to the south at 18^h and 4·7° to the north at 6^h. A daily variation of this kind seems to have been observed.

Alfvén also discusses the variation with local time of the direction of the auroral arcs and the current system due to the streaming of the electrons on the morning side of the earth.

He estimates (1939, p. 204) that the potential difference across the forbidden region (from east to west) is of the order of 10^4–10^5 volts, so that electrons impinging on the upper atmosphere would have the energy required for penetration down to the auroral levels.

Quantitative tests of the theory can be made by deduction of (i) the polar distance of the auroral zone, (ii) the energy of the auroral electrons, (iii) the time of passage of the stream from the sun to the earth, (iv) the total current required to explain the storm. These depend on two adjustable parameters, corresponding roughly to the energy of the particles in the stream and the value of the solar electric field E. By a suitable choice of these, Alfvén is able to get satisfactory values for each of the four quantities.

Fig. 18



The auroral zone.

§ 34. The Electric Current System

We have already seen that, in Alfvén's theory, when the stream invades the earth's magnetic field it is flattened into a thin disc in the magnetic equatorial plane. The penetration of the stream is, however, limited by the boundary of the forbidden region R.

The inhomogeneity drift of the charges in the earth's field is eastwards for the electrons and westwards for the ions and the greater part of the equatorial disturbance of a geomagnetic storm is attributed by Alfvén to the drift of the ions and electrons outside this forbidden region.

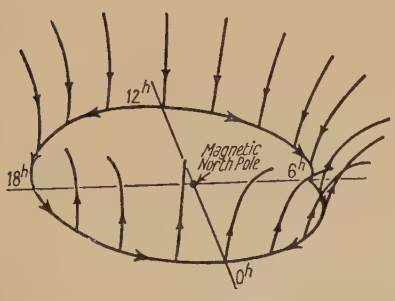
In the equatorial plane the current flow along concentric circular arcs since the cross-drift is the same for both positive ions and electrons ; they intersect the boundary of R when their radii are smaller than the maximum distance of R from the earth (fig. 19). This entails the transport of charges to the boundary of R and Alfvén supposes that this is then discharged to the polar region along the auroral zones. Thus the charges

leave the equatoral plane and their flow is continued along the magnetic lines of force until they reach the polar regions along the auroral zone. Along this zone, however, the electrical conductivity of the atmosphere is supposed sufficient to allow the discharge currents to continue their flow along the zone (fig. 20). This current system is responsible for the

Fig. 19

Resultant current system.

Fig. 20



Current system of the auroral zone.

polar disturbance and completes the circuit of the currents in the equatorial plane which, according to Alfvén, is responsible for the equatorial disturbance. It will be seen that the current along the auroral zone is westward from 0^h to 12^h and eastward from 12^h to 18^h, the former being the more intense.

Alfvén has investigated the characteristics of the magnetic field which such a system of currents would produce by means of a model and has compared it with the form of the S_D field of an average magnetic storm. He finds good qualitative agreement ;* nevertheless, no direct attempt is made to estimate the intensity of the current. This depends essentially on the particle density in the stream, since the current is due to the inhomogeneity drift and the cross-drift u_E is the same for the positive ions and electrons.

§ 35. The Particle Density in the Stream

In his first paper (1939) Alfvén estimated the intensity of the electric field in the stream by considering the rate of escape of charges from the flanks of the stream, and to make the time of passage from the sun to the earth of the order of one day, concluded that the particle density in the stream near the earth is of the order of 10^{-4} particles/c.c.—an astonishingly low value. It seems unlikely that the density can be as low as this for the following reasons.

The current in the stream is due to the inhomogeneity drift of the ions and electrons since the cross-drift is the same for the positive ions and electrons and we may neglect the smaller drift of the ions. By (6) the inhomogeneity drift of electron is of the order of $\frac{T}{eH^2}\frac{\partial H}{\partial r}$, where T is the kinetic energy of the particle. Denoting by N the electronic density, expressing T in electron volts, V, and using the fact that $T \propto H$ we find the current density to be

$$j = \frac{N\mu}{H}\frac{\partial H}{\partial r},$$

when $\mu = eV/H = \text{constant}$. Further $\partial H/\partial r \sim H/r$, and thus $j \sim N\mu/r$. If $\mathbf{H}'$ is the magnetic field produced by the current, the circuital relation

$$\text{curl}\,\mathbf{H}' = 4\pi\mathbf{j}$$

gives, as far as orders of magnitude are concerned, $H' \sim 4\pi jD$, where D is comparable with the thickness of the equatorial ring current. Combining this with the above estimate of j, we have $H' \sim 4\pi N\mu D/r$, or $N \sim rH'/(4\pi\mu D)$. Now $\mu = eV/H$ and is constant : according to Alfvén the energy of the electrons at emission is of the order of 10^8 ev and taking the value of the solar magnetic field to be of the order of 50 gauss, we find $\mu = 3 \times 10^{-6}$. Also, the stream near the earth becomes

* Kirkpatrick however questions the agreement in a recent paper (*J. Geophys. Res.*, 57 (1952), 511).

flattened in the equatorial plane so that we may expect that the thickness of the equatorial current D to be small compared with its mean radius r. Thus $r/D \gg 1$, and $N \gg H'(4\pi\mu)^{-1}$.

If the equatorial current is responsible for the equatorial disturbance of a geomagnetic storm, as Alfvén suggests, H' must be comparable with the disturbance of an average storm, about 50 γ or 5×10^{-4} gauss, say; the above inequality then implies that $N \gg 10$ particles per c.c. Thus, the density of the stream cannot be as low as Alfvén estimates.

§ 36. Some Unsatisfactory Features of the Theory

Cowling (1942) has shown that Alfvén's theory is open to a number of criticisms.

(1) If the ions entering the forbidden region R along CDC′ (see fig. 18) are immediately repelled to the poles along the lines of force, electrons must also be immediately discharged to the poles as soon as they cross the boundary C′F′ of the earth's shadow where they become separated from the ions. The auroral zone is thus the projection of C′F′ along the lines of force, and not the projection of C′D′C′ and its shape is altogether different from Alfvén's curve (fig. 18).

(2) Cowling also draws attention to a serious energy difficulty. Near the earth the inhomogeneity drift of the ions is small compared with the cross drift and since this is perpendicular to the electric field an ion drifts roughly along a line of constant electric potential. During the discharge of the positive ion to an auroral zone, the transverse component of velocity v_T varies according to (10). Since H increases about 1000-fold between CDC′ (fig. 17) and the earth's atmosphere, an ion reaches the earth with very much greater kinetic energy than it had before reaching CDC′. This is only possible if its final electric potential energy is smaller than its initial value. Whilst this may be true of the ions leaving the boundary CDC′ in the vicinity of C′, on account of the electric field parallel to C′C, it cannot be true for the majority of the ions. Cowling concludes "that diamagnetic repulsion prevents the particles from reaching the earth's atmosphere, and that electrostatic forces prevent more than a slight separation of ions from electrons. It seems that a stream of ionized particles can reach the earth only if either the particles possess such large energies that we cannot divide their motion roughly into a spiral motion along a line of force and a relatively slow drift across the lines or force, or if the stream is so dense that the leading particles are able to shield the body of the stream from diamagnetic repulsion".

Meinel's recent observations of an emission of strongly shifted hydrogen lines in the aurorae spectrum, suggesting the injection of high speed protons into the earth's atmosphere, is unfavourable to the production of aurorae by electrons on the night side of the earth, as in Alfvén's theory.

No explanation is given of the first phase of a magnetic storm, and to explain the main phase of a magnetic storm the particle density must considerably exceed his estimate of 10^{-4} particles per c.c. and be at least of order unity (§ 35).

(*c*) Martyn's Theory

§ 37. The Chapman–Ferraro Theory of Magnetic Storms

Chapman and Ferraro (1931, 1932, 1933, 1940) have developed a theory of magnetic storms based on the emission of a neutral ionized stream from the sun considered in §§ 20–26. Only the theory of the first phase has been developed to any extent and their treatment of the main phase is at present only tentative. They made no attempt to explain aurorae, though they believed that, if correct, the theory would provide a starting point for an auroral theory; a first essay in this direction has been made by Martyn.

A brief description of the theory is necessary beforehand. The speed of the particles is assumed to be of the order of 1000 km/sec, because of the lag of about one day between great solar flares and the onset of the magnetic storms associated with them. No attempt was made to formulate a theory of emission and no account was taken of the powerful sunspot magnetic fields, or of a possible general magnetic field. Chapman and Ferraro considered that such a stream would be able to move freely through a strong solar magnetic field on account of the polarization electric field set up which overcomes the tendency of the ions and electrons to spiral round the lines of force.

Chapman (1948) estimates this field to be of the order of 25 volts/cm as against 3300 volts/cm for Alfvén's field.

An ionized gas is a good electrical conductor and on arrival near the earth electric currents are induced in the surface layers of the stream. These shield the interior of the stream from the influence of the magnetic field, as in the case of a perfect conductor, so that the particles therein are undeflected; these overtake the particles in the surface layers to form a new surface and the stream particles can thus advance further into the earth's field than would be the case if they were alone in the field. The action of the earth's magnetic field on the induced surface current retards the motion of the surface, the retardation being greater for the points nearer to the earth. This action results in the carving out of a hollow in the stream which deepens as the stream advances (fig. 21).* Beyond the earth the magnetic field polarizes the stream and surface charges are produced over the walls of the hollow, positive on the morning side of the hollow and negative on the evening side (see fig. 22).

After a time the front surface of the stream is brought to rest and the distance of the nearest point from the earth can be estimated very simply by equating the rate at which momentum of the solar stream is destroyed to the normal pressure exerted by the magnetic lines of force

* The fact that the surface currents shield the stream from the earth's magnetic field outside can be interpreted as implying that the tubes of force are unable to penetrate the stream surface and are thus pushed forward by the stream and crowded together in the hollow space. The magnetic force in the hollow is thereby increased.

on the stream surface, as was first suggested by Martyn. If N, m, u denote the number density, mass and velocity of the stream and H the magnetic field, this gives

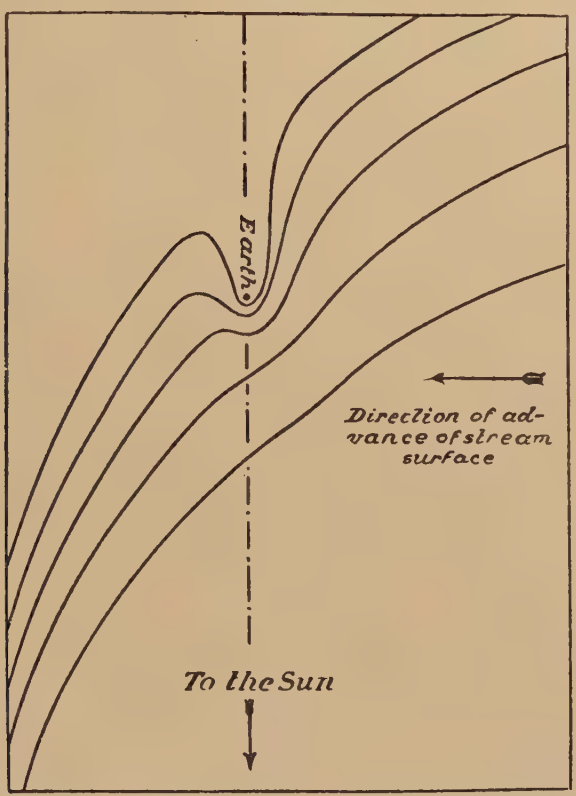
$$\tfrac{1}{2}Nmu^2=H^2/8\pi. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

In the geomagnetic equatorial plane $H=0{\cdot}33z^{-3}$, where z is the distance from the earth's centre, measured in earth radii. Assuming that $u=10^8$ cm/sec and that the stream consists of protons and electrons, (16) gives

$$z=8{\cdot}8N^{-1/6}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

Thus the size of the hollow varies little over wide ranges of N.

Fig. 21



Sections by the earth's equatorial plane of the advancing front surface of a corpuscular stream, illustrating diagrammatically the formation of a hollow space round the earth by the action of the geomagnetic field on the neutral ionized stream.

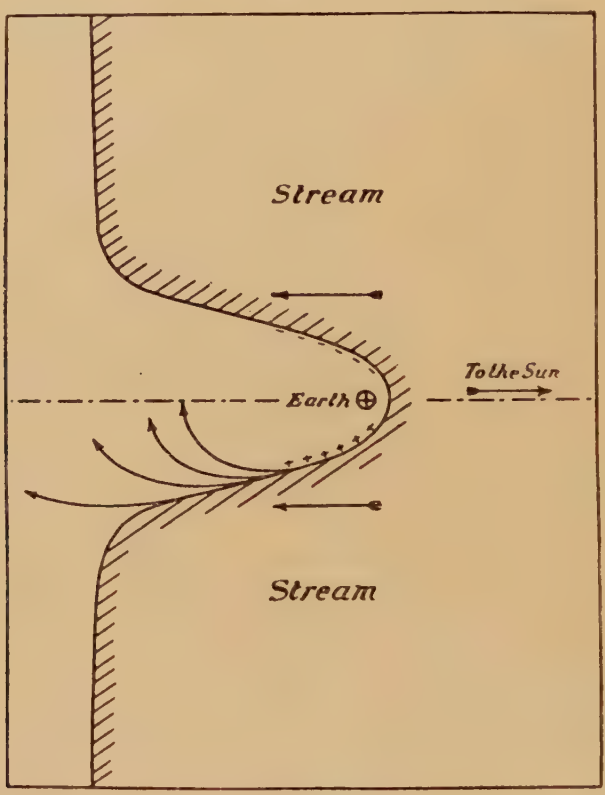
Chapman and Ferraro showed that the increase in the magnetic field of the induced currents outside the stream could be approximately estimated by replacing the currents by an image dipole at a distance $2z$

from the earth. To explain the rise of about $20\,\gamma$ in the earth's magnetic field during the first phase of a magnetic storm, the point of the stream must reach to within a distance z from the earth given by

$$0{\cdot}33/(2z)^3 = 25\times 10^{-5},$$

giving $z=5{\cdot}5$. Hence, by (16), N is of the order of 20 protons per c.c. or about 1 in 4 c.c. if the ions in the stream are Calcium ions. The energy of the electrons and protons at a speed of 10^8 per sec are respectively 2·5 ev and 4600 ev as against the 10^8 ev of the electrons in Alfvén's theory.

Fig. 22



Illustrating diagrammatically, in an equatorial section, the sign of the electric charges on the surface of the hollow carved in the advancing neutral ionized stream by the geomagnetic field. The curved arrows indicate the curved paths described, under the deflecting influence of the field, by charges which escape from the surface.

Once the hollow becomes stationary relative to axes fixed in the earth, the positive surface charges on the walls tend to bridge the gap as indicated in fig. 22 ; in this they will be helped by the electric field across the gap, but because of the earth's magnetic field they can only do so at certain distances from the earth. Chapman and Ferraro considered this to be

the beginning of the main phase of a magnetic storm and of the formation of the westward ring current, but the suggestion is only tentative and no attempt has been made to discuss the growth of the ring quantitatively. They considered the equilibrium and stability of the ring current once formed on the hypothesis that the ring current is polarized radially, the positive surface charge being on the inner surface of the ring, so that the electrons and ions describe nearly circular paths round the earth instead of spiralling in the equatorial plane with a drift velocity (inhomogeneity drift) as in Alfvén's theory. Mechanical considerations show that only a ring carrying a westward current is possible. For if u_i and u_e denote the speeds of the ions and electrons taken positive westwards, m_i and m_e their masses, and E the radial electric field, the equations of steady motion of ion and electron are respectively,

$$-m_i u_i^2/r = eE - eu_i H, \quad \text{e.m.u.}$$
$$-m_e u_e^2/r = -eE + eu_e H.$$

Adding and neglecting the smaller kinetic energy of the electrons we have approximately

$$m_i u_i^2 = reH(u_i - u_e). \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (18)$$

Thus $u_i > u_e$, so that the current is westwards.

The magnitude of the current will be

$$i = Ne(u_i - u_e)A, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

where A is the area of cross section of the ring. This current produces a decrease of the earth's field at its surface of amount

$$\Delta H = 2\pi i/r\,; \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (20)$$

combining (17)–(20), we find

$$A = 2z^5 a^2 \Delta H/0{\cdot}33,$$

where a is the radius of the earth. Taking $z = 5{\cdot}5$ and $\Delta H = 50\gamma$, the observed decrease in H in the main phase of an average magnetic storm, we have

$$A = 15a^2.$$

If we assume the cross section to be circular for simplicity, and of radius b, we have $A = \pi b^2$, so that $b = 2a$ approximately or about equal to the diameter of the earth.

Chapman and Ferraro also discussed the stability of the ring current and showed that because of its large electromagnetic momentum, the decay of the current through resistance would be extremely slow. This is in accord with characteristic slow decay of the main phase of a magnetic storm. Martyn has suggested that because of the electromagnetic pressures in the stream, due to the interaction between the electric currents in the stream and the earth's magnetic field, the hollow in the stream will close up, the flow resembling that of a liquid past a submerged obstacle. In this respect he departs from the Chapman–Ferraro theory and suggests that the ring-current is formed by forces which constrain

the stream to flow in curved paths round the earth. The following considerations do not seem to lend support to this views. They were advanced by Chapman in 1923 in a first discussion of the steady motion of a neutral ionized stream in the earth's magnetic field, in which only the polarization of the stream was taken into account; as Chapman showed, the stream is unlikely to be bent round the earth appreciably except close to the earth's surface. We can readily verify this by estimating the curvature of a stream line as follows: let **P** denote the polarization of the stream at any point, and **E** and **H** the electric and magnetic fields. Then, as Chapman showed, the force on an element of volume, per unit volume, is

$$(\mathbf{P}\,.\,\mathrm{grad})\mathbf{E}+\{(\mathbf{V}\,.\,\mathrm{grad})\mathbf{P}\}\wedge\mathbf{H},$$

V being the velocity of the element. The first term arises from the action of the electric field **E** on the Poisson volume distribution of charge, $-\mathrm{div}\,\mathbf{P}$, and surface distribution, P_n, over the surface of the element. The second term arises from the action of the magnetic field on the polarization current of density $d\mathbf{P}/dt$, or $(\mathbf{V}\,.\,\mathrm{grad})\mathbf{P}$ for steady motion. Thus, the equation of motion of the element is*

$$N m\dot{\mathbf{V}}=(\mathbf{P}\,.\,\mathrm{grad})\mathbf{E}+\{(\mathbf{V}\,.\,\mathrm{grad})\mathbf{P}\}\wedge\mathbf{H}. \quad . \quad . \quad . \quad . \quad . \quad (21)$$

The electric field **E** acting on the moving element is, ignoring relativity corrections, $(1/c)\mathbf{V}\wedge\mathbf{H}$ very nearly. To relate **P** and **E** we note that the stream is a good conductor so that, as far as order of magnitude are concerned, $\mathbf{P}=\mathbf{E}/4\pi$.

Let ρ be the radius of curvature of a stream line at a distance r from the centre of the earth, then, considering order of magnitude only in which we replace $\partial E/\partial r$ by E/r, etc., the equation of normal acceleration is, by (21)

$$N m V^2/\rho \sim V^2H^2/(4\pi c^2 r),$$

or

$$\rho/r \sim 4\pi N m c^2/H^2\,; \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

this relation is independent of V since both the normal acceleration and deflecting force are proportional to V^2. It agrees, as far as orders of magnitude are concerned, with Chapman's more detailed discussion. If we use the relation (16), giving the distance from the earth at which the stream is brought to rest, (22) becomes

$$\rho/r \sim c^2/V^2,$$

so that unless V approaches the speed of light, or N is considerably smaller than the value given by (16), ρ is large compared with r. Thus the curvature of the stream lines which, as Chapman showed are concave to the earth, is negligible and as Chapman had already inferred in his 1923 paper, the ions and electrons are able to move along nearly straight paths with little or no bending of stream round the earth.

---

* The remaining symbols have the same meaning as before.

§ 38. Martyn's Theory of the Aurora

The polarization electric field, acting radially (which enables the ions and electrons to flow round the earth in circular paths of larger radius then their respective spiral radius) induces charges on the surface of the ring current. These charges are unstable because of the mutual repulsion of the charge in the layers and are repelled along the magnetic lines of force, the leakage being made good from the body of the stream to maintain the electric field.

Martyn drew attention to the important fact that this leakage may be more than a secondary phenomenon as far as aurorae and magnetic storms are concerned.* The polarization electric field which nearly balances the Lorentz force $\mathbf{V} \wedge \mathbf{H}/c$ (e.s.u.) on a particle in the ring, is of the order of 10^{-3} volts/cm and extends over a distance of the order of 10^9 cm. The gain of potential energy of particles accelerated by this field from the surface of the ring to the earth's atmosphere is of the order of 10^6 volts ; the particles would thus acquire speeds of the order of 10^9 cm/sec which would be adequate to enable them to penetrate to the auroral levels (see table 2).

As a first approximation we may consider the projection along the lines of force of the inner and outer boundary of the ring current in the equatorial plane over the surface of the earth to mark the boundaries of the region of precipitation of the charges. If the projection of the central filament of the ring current intersects the surface in colatitude θ_1 we have

$$\operatorname{cosec}^2 \theta_1 = z,$$

where z is the radius of this filament in earth radii. For a ring of mean radius 5·5, corresponding to the distance from the earth at which the front of the stream is brought to rest, we find $\theta_1 = 25°$. Further, if $r(=2b)$ be the radial separation of the lines of force which intersect the inner and outer boundaries of the ring, we have, from the equation of a line of force, $\sin^2 \theta / r = \text{constant}$,

$$\delta r = -\frac{2a \cos \theta_1}{\sin^3 \theta} \delta\theta.$$

Taking $\delta r = 2{\cdot}6 \times 10^9$ cm, as before, we obtain $\delta\theta = 6°$ for the width of the region. Martyn's identifies this zone of precipitation of the charges escaping from the ring with the auroral zone, and points out that the width of 6° agrees well with the estimate derived by Chapman and Vestine (1938). A more accurate estimation of the angular radius of the auroral zone and the width to be expected has been made by Nagata (1952) who takes into account also the magnetic field of the ring-current. The following table, due to Nagata, shows the variation of the inner and outer angular radii of the zone, θ_i and θ_e, for rings of radius za producing a decrease of the geomagnetic field ΔH at the earth's surface.

* Chapman and Ferraro had suggested this possibility though they gave no detailed discussion (see Chapman (1948), p. 131).

It will be seen that for a ring current producing a moderate disturbance (say 20γ) the results are in good agreement with Martyn's approximate theory. If the current in the ring is large the surface charges may be trapped by the magnetic field of the ring and cannot then escape. This is indicated by the absence of entries in the table in the columns for $z=8$ and 10. Two interesting points are brought out by this table ; firstly the auroral zone is shifted southwards (in the northern hemisphere) as the magnetic disturbance produced by the ring increases, in accordance with observed facts. Secondly, we must expect the position of the ring to be nearer the earth at times of large magnetic disturbance ; in fact if the rings are further out than a distance of 10 earth radii, the auroral zone could not extend further south than about 23°. The fact that the aurora is seen in geomagnetic latitudes as low as 35° indicates that on Martyn's theory, we must expect that at times of great magnetic storms the ring current to be nearer to the earth. This would be expected on the Chapman–Ferraro theory since the greater the penetration of the stream surface into the earth's magnetic field the greater the magnetic disturbance. On the Chapman–Ferraro theory the ring current must be symmetrical with respect to the geomagnetic axis. This implies that the auroral zone is very nearly circular, in contrast to the asymmetrical shape of the auroral zone on Alfvén's theory.

Table 4

ΔH	$z=6$		$z=8$		$z=10$	
	θ_e	θ_i	θ_e	θ_i	θ_e	θ_i
0γ	26° 34′	22° 12′	22° 12′	19° 28′	19° 28′	17° 33′
10	27 00	22 51	23 10	20 52	21 29	22 02
20	27 26	23 29	24 13	22 11	22 28	22 16
40	28 19	24 41	26 05	24 38	—	—
60	29 07	25 51	27 24	26 56	—	—
80	29 56	26 59	—	—	—	—
100	30 43	28 11	—	—	—	—

§ 39. The Auroral Zone Current

The charges escaping from the charged layers on the walls of the hollow along the lines of magnetic force will, after a time, render these lines of equipotentials. Thus the polarization electric field will be transferred to the auroral zone and there give rise to a meridional electric field of the same order of magnitude as in the charged layers. The incoming particle will produce, moreover, enhanced ionization and conductivity in the ionosphere, especially in the lower levels of penetration : such enhanced ionization at times of auroral displays has indeed been found by radio-sonde (Ratcliffe and White 1933, Lovell *et al.* 1947). Thus electric currents will flow along meridians across the auroral zone. The current

system envisaged by Martyn is one flowing in a closed system of sheets from the stream to the zone and thence back to the stream. The currents in the sheets, being opposite on the two opposite sides, will only produce a small effect at the earth's surface, but may be intense enough to contribute to the polar magnetic disturbance. Martyn estimates that the potential difference between the inner and outer boundaries of the zone may well exceed 10^5 volts, so that the meridional electric field E_a is rather more than 10^{-3} volts/cm. In conjunction with the nearly vertical electric field, H_a, in the auroral zone, this electric field will produce a drift motion of the ions of both signs of magnitude

$$\left(\frac{E_a}{H_a}\right)\frac{\omega^2}{\omega^2+\nu^2},$$

where $\omega(=eH/m)$ is the gyromagnetic frequency of the charge and ν the ionic collision frequency.

This drift will be westerly on the morning side and easterly on the evening side of the earth. As the flow of the stream from the sun diminishes and the momentum of the ring dominates the circulation of the ring, the drift will be westerly in all parts of the zone, since the electric field will then be due to the charges escaping from the ring.

If the masses and collision time-interval of the ions and electrons were identical, these drifts would, during the first day of the storm, merely transfer ions from the sunlit to the dark side of the auroral zone. It is highly improbable that both the masses and collision time-interval of both the ions and electrons are identical. Martyn states that the magnetic and ionospheric effects can be readily explained if the ions drift somewhat faster, say 10 to 30% faster, than the electrons. In this case, a Hall current amounting to about half-a-million amperes will flow round the auroral zone in the first day of the storm westward on the dawn hemisphere eastward on the sunset meridian. This Hall current sets up a polarization field tending to impede the flow, which is positive on the midnight, and negative on the noon meridian. The potential of this field in the auroral zone can be represented by $S_0 \cos\psi$, where ψ is the longitude measured from the midnight meridian. Martyn finds that this potential distribution sets up an ionospheric current system in all parts of the world which has the strength and form required to explain the main features of the S_D field, as described by Chapman and Vestivie (1938).

§ 40. Discussion of Auroral Theories

Of the various interesting theories of the aurora which have been proposed none can be said to be complete or satisfactory. The theories of Störmer and Birkeland reproduce many analogies with the observed forms of the aurora and their geographical distribution and diurnal variation. But, as wi h all one-sign theories, the criticism first directed against them by Schuster that the stream could not hold together against the electrostatic repulsion of its parts is really destructive.

Likewise, the neglect of the powerful local electrostatic fields in Alfvén's theory, once the ions invade the forbidden region of the electrons, to which Cowling drew attention, shows that the theory remains in an unsatisfactory state. The theory is moreover largely qualitative and where quantitative arguments are given, these are usually unconvincing. Chapman (1948) states that before the stream reaches a distance of 10 solar radii from the sun, the lateral drift would have completely separated the positive and negative parts of the stream. Alfvén's theory has nevertheless many original and attractive features, some in common with Martyn's theory.

As regards the Chapman–Ferraro theory, only the theory of the first phase is at all complete and their theory of the main phase is at present only a qualitative sketch. Martyn has made a valuable contribution to their theory by pointing out that the charges escaping from the ring current would acquire energy sufficient to give the observed penetration of aurorae. He suggested that the hollow in the stream would close up on the night side of the earth, on account of maxwellian pressures on the walls of the hollow: but he gives no quantitative discussion and for reasons given in § 37, the writer is inclined to doubt whether the hollow does in fact close up. Martyn's theory is perhaps the more promising one at present. It has several features in common with Alfvén's: both suppose that the equatorial ring current flows at distances of a few earth radii from the earth, and in both theories the charged particles are supposed to be accelerated near the earth by electric fields so as to acquire sufficiently high energies to enable them to penetrate down to auroral levels. Both identify the auroral zone with the projection along the lines of force over the earth's surface of the boundary of the forbidden region (in the case of Alfvén's theory) or of the ring current (in the case of Martyn's theory). Their theories differ in that in Alfvén's theory the positive ions are precipitated over the morning side of the earth and the electrons over the night side, whilst in Martyn's theory both positive ions and electrons are precipitated over the 'auroral zone'. In this connection Meinel's observations of the Doppler shift of the $H\alpha$ line in the auroral spectrum indicating the entry of protons in the earth's atmosphere, are unfavourable to a theory in which electrons alone are precipitated over the night side of the earth.

The low particle density required by Alfvén (10^{-4} particles per c.c.) also contrasts with the density required by the Chapman–Ferraro–Martyn theory (about 20 protons or 1 calcium atom per c.c.). The writer believes that a stream whose particle density is as low as Alfvén supposes is unlikely to be able to account for geomagnetic disturbance (§ 35).

The current system, to which the polar magnetic disturbance is attributed in both theories, is supposed to flow mainly in the auroral zones, westward on the morning side, eastward on the evening side, as suggested by Chapman's analysis of the S_D field. But the origin of these currents differs in the two theories. In Alfvén's theory there is a poleward current from the boundary of the forbidden region facing the sunlit

hemisphere ; and the circuit is completed by a return current via the auroral zone to the boundary of the forbidden region facing the nightside hemisphere. In Martyn's theory the poleward and return currents flow in neighbouring sheets and the auroral zone current is due to a differential drift velocity of the ions and electrons.

Alfvén has sought to verify his theory by a laboratory experiment. This has been carried out by Malmfors (1946) and resembles Birkeland's experiment with a terrella except that an electric field normal to the direction of magnetization of the terrella is applied and the ionization is started by an electron gun. As in Birkeland's experiments, the ions and electrons are deflected poleward by the (dipole) magnetic field of the terrella and there produce two luminous rings, one around each pole. The rings are, moreover, eccentric, in accord with Alfvén's theory. Alfvén maintains that this experiment is important in showing the essential part played by the solar magnetic field. But such an experiment can only have direct bearing on the actual natural motions of the charges if the reduction in linear scale corresponds exactly with the conditions obtaining in the natural case. This condition does not appear to obtain in Malmfors's experiment.

Only a brief mention can be made of the 'ultraviolet light' theory of Maris and Hulburt (1928, 1929, 1930). This, like Dauvillier's theory (1932), ascribes the production of the aurora to terrestrial particles though in every other respects differs greatly from it. The authors suppose that in the exosphere, the fringe of the upper atmosphere where collisions are rare, some of the molecules may, by superelastic collisions, be driven upwards to heights of 30 to 40 thousand kilometers above the earth. An aurora or magnetic storm is supposed to occur when the sun emits an abnormally intense ' blast ' of ultraviolet light, thereby ionizing these high flying molecules. The ions and electrons so liberated are at once subject to the guidance of the earth's magnetic field, which compels them to travel along the magnetic lines of force. The lines of force that leave the earth's surface along the auroral zone meet the equator at a height of 30 000 to 40 000 km above the surface, and it is for this reason that Hulburt and Maris suppose that the auroral particles originate at this level.

The blast of ultraviolet light is supposed to facilitate further the production of high speed particles by exciting more particles at lower levels. The particles take about 3 hours to attain their supposed high level ; some particles will be ionized before reaching their topmost level, will enter the atmosphere in lower latitudes than usual—in accordance with the observed increase in the auroral radius at times of magnetic storms.

As a theory of aurorae, this has been criticized by Chapman on the grounds that the velocity of free fall under gravity acquired by the particles on arrival in the earth's atmosphere from 40 000 km is about 10 km/sec, which is far too small to account for penetration down to auroral levels. The theory has been criticized as a theory of magnetic storms by Chapman (1930) and McNish (1937).

§ 41. Conclusion

The diversities of auroral theories so far proposed may indeed seem perplexing. The difficulties encountered by corpuscular theories of one sign are insuperable, and the most promising line of attack now seems to be the hypothesis that aurorae are due to phenomena accompanying the advance of a neutral ionized gas in the earth's magnetic field, as in the theories of Alfvén and Martyn.

Alfvén's theory requires the additional hypothesis that the sun should possess a general magnetic field but, as was mentioned in § 32, there appears to be little evidence to support this. Moreover, the theory offers no explanation of the first phase of magnetic storms and, as Cowling has shown, it is open to a number of objections.

There remains Martyn's theory based on the theory of magnetic storms developed by Chapman and Ferraro. In some respect this theory remains the most promising starting point for an auroral theory and the authors feel sufficiently confident that it is substantially correct. Whilst Martyn's theory is an interesting and valuable contribution to the subject it would be premature to comment on the correctness of his views : there is a great need for caution in a subject which abounds in pitfalls. This is in part due to our lack of understanding of the phenomena accompanying the motion of a neutral ionized gas in a magnetic field ; and this subject would certainly repay further study just as further observation of the auroral spectrum, especially at times of intense magnetic storms, would add to our understanding of the problem.

References

Abetti, G., 1929, *Second Report on Solar and Terrestrial Relationships*, p. 9.

Alfvén, H., 1939, *K. Svenska Vet. Akad. Handl.* (3) **18,** No. 3 ; 1940, *Ibid.*, **18,** No. 9 ; 1950, *Cosmical Electrodynamics* (Oxford), p. 181.

Allen, C. W., 1938, *Observatory*, **61,** 136.

Babcock, H. D., 1923, *Astrophys. Si.*, **57,** 209.

Bartels, J., 1932, *Terr. Mag. Atmos. Elec.*, **37,** 1.

Bates, D. R., 1949, *M.N.R.A.S.*, **109,** 215.

Bernard, R., 1948, *Gassiot Committee Conference* (London : Physical Society) p. 91.

Birkeland, K., 1896, *Arch. Sci. Phys. Genève*, **4,** 497.

Boller, W., 1898, *Gerlands Beitr. zur Geophysik*, **3,** 56, 550 ; 1902, *Catalog der in Norwegen bis Juni* 1878 *beobachteten Nordlichter* (Kristiàna).

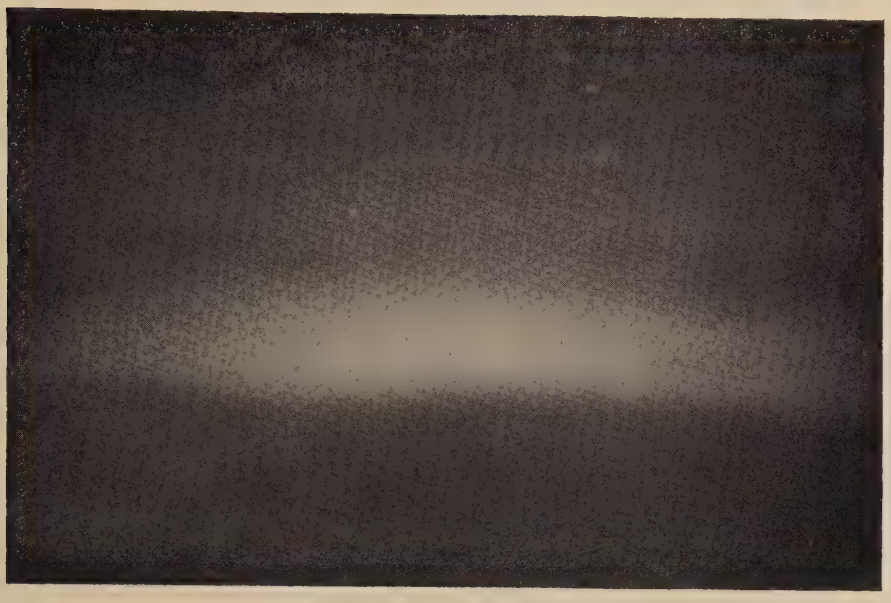
Brüche, E., 1931, *Terr. Mag. Atmos. Elect.*, **36,** 41.

Brück, H. A., and Rutlantt, F., 1946, *M.N.R.A.S.*, **106,** 130.

Cavendish, H., 1790, *Phil. Trans. Roy. Soc.*, **80,** 101 ; or see Papers **2** (Cambridge 1921).

Chapman, S., 1923, *Proc. Camb. Phil. Soc.*, **21,** 577 ; 1929, *M.N.R.A.S.*, **89,** 456 ; 1930, *Ibid., Geophys. Suppt.*, **2,** 296 ; 1931, *Nature, Lond.*, **127,** 341 ; 1932, *Ibid.*, **130,** 764 ; 1937, *Phil. Mag.*, **23,** 657 ; 1948 a, *Ann. de Geophys.* (Lyon Report) ; 1948 b, *Gassiot Committee Report* (London : Physical Society), p. 120.

Chapman, S., and Ferraro, V. C. A., 1929, *M.N.R.A.S.*, **89,** 470 ; 1931, *Terr. Mag. Atmos. Elec.*, **36,** 77, 171 ; 1932, *Ibid.*, **37,** 147, 421 ; 1933, *Ibid.*, **38,** 79 ; 1940, *Ibid.*, **45,** 245.



(*a*) Diffuse Homogeneous arc *HA*.

(*b*) Diffuse *HA* with sunlit ray *and* noctilucent cloud along horizon.
0006^{h} G.M.T. 25 July, 1950.

(*a*) Homogeneous arc *HA* changing into Homogeneous band *HB* (thin *RB*).

(*b*) Rayed Arc.

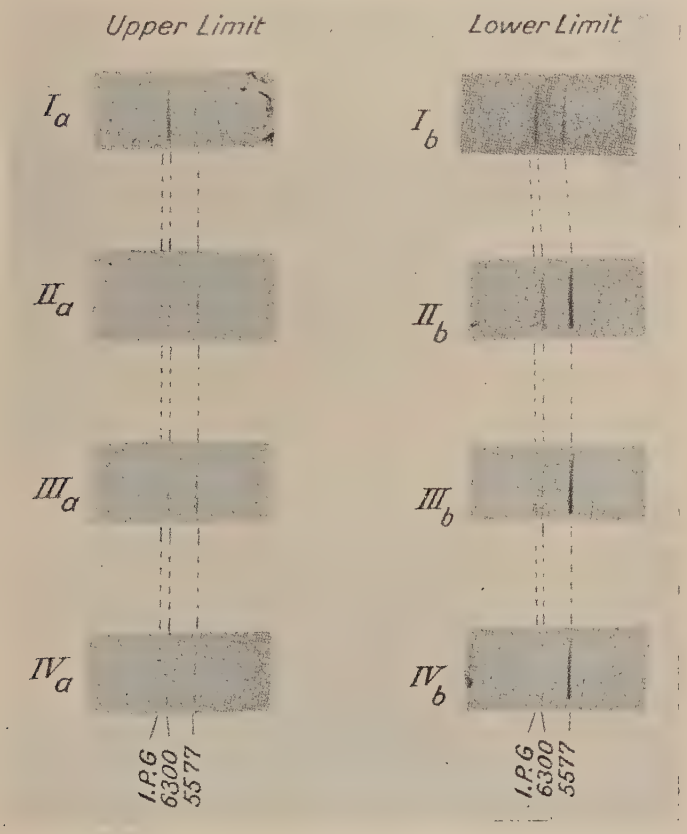
(*a*) Eastern end of rayed arc developing to curtain.

(*b*) Corona formed by draperies.

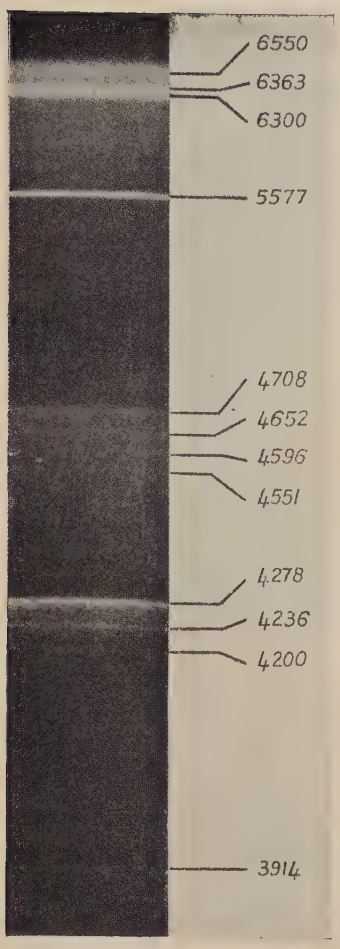
(*a*) Corona of rays.

(*b*) Ray bundle.

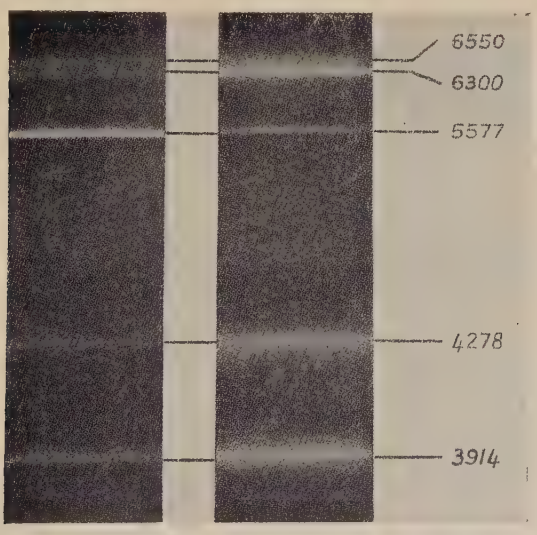
(*a*) Two auroral spectrograms taken by Vegard and Tönsberg, *A* with effective exposure 37 hours (from 1935 October 15 to 1936 March 28), and *B* with effective exposure 15 hours (from 1936 December 3 to 1937 April 10).



(*b*) Spectrograms showing the enhancement of the red *OI*-line (6300) relative to the green *OI*-line (5577), with increasing altitude ; I, 1937 October 11, II–IV, 1937 October 12 (Vegard, *Geof. Publ.*, xii, No. 5). *I.P.G.*=1st positive bands of nitrogen.



(*a*) Spectrum of the summits of blue auroral rays at a height of about 400–650 km taken at Oslo by Störmer, 1938 September 15.



(*b*) The same spectrum (on the left) compared with a spectrum of the light from yellow-green auroral curtains, height of lower border about 92 km.

Chapman, S., and Milne, E. A., 1920, *Quart. J. Roy. Meteorol. Soc.*, **46,** 384.
Chapman, S., and Vestine, E. H., 1938, *Terr. Mag. Atmos. Elec.*, **43,** 351.
Cowling, T. E., 1942, *Terr. Mag. Atmos. Elec.*, **47,** 209.
das Gupta, N. N., and Ghosh, S. K., 1946, *Rev. Mod. Phys.*, **18,** 225.
Dauvillier, A., 1932, *Rev. gén. elect.*, **31,** 443.
Davies, F. T., 1930, *Terr. Mag. Atmos. elec.*, **36,** 199.
de Mairan, M., 1773, *Traité physique et historique de l'aurore boréale* (Paris).
Dufay, J., and Gauzit, J., 1938, *Comptes Rendus Acad. Sc. Paris*, **206,** 619.
Dufay, J., Gauzit, J., and Tscheng-Mao-Lin, 1941, *Cahiers de Physique*, **1,** 1.
Dufay, J., and Tscheng-Mao-Lin, 1942, *Cahiers de Physique*, **8,** 51.
Ferraro, V. C .A., 1930, *M.N.R.A.S.*, **91,** 174.
Fritz, H., 1873, *Verzeichnis beobachteter Polar Lichte* (Wien); 1881, *Das Polarlicht* (Leipzig).
Gartlein, C. W., 1951, *Phys. Rev.*, **81,** 463.
Goldstein, S., 1881, *Wiedemanns Ann.*, **12,** 266.
Götz, F. W. P., 1947, *Experentia*, **3,** 185.
Greaves, W. M. H., and Newton, H. W., 1928, *M.N.R.A.S.*, **88,** 556; 1929, *Ibid.*, **89,** 84.
Gnevishev, M. N., and Ol, A. I., 1946, *Terr. Mag. Atmos. Elec.*, **51,** 163.
Harang, L., 1950, *Intern. Astrophys. Series* (London).
Harang, L., and Bauer, W., 1932, *Gerlands Beiträge zur Geophysik*, **37,** 109.
Hulburt, E. O., 1928, *Phys. Rev.*, **31,** 1038; 1929, *Ibid.*, **34,** 344; 1930, *Ibid.*, **36,** 1560; 1931, *Terr. Mag. Atmos. Elec.*, **36,** 23.
Kahn, F. D., 1949, *M.N.R.A.S.*, **109,** 324.
Kaplan, J., 1934, *Phys. Rev.*, **45,** 671; 1936, *Comm. Relation Solares Tenestres*, 4th Report, pp. 92–7.
Kiepenheuer, K. O., 1952, *J. Geophys. Res.*, **57,** 113.
Lee, A. W., 1930, *Meteorol. Office, Profess. Notes.* No. 56 (London).
Lindemann, F. A. (Lord Cherwell), 1919, *Phil. Mag.*, **38,** 669.
Lord Rayleigh, 1928 a, *Proc. Roy. Soc.* A, **119,** 11; 1928 b, *Nature, Lond.*, **122,** 315; 1930, *Proc. Roy. Soc.* A, **129,** 458.
Lovell, A. C. B., Clegg, J. A., and Ellyett, C. D., 1947, *Nature, Lond.*, **160,** 372.
Lovering, J., 1868, *Mem. Amer. Acad., New Series*, **10.**
Malmfors, K. G., 1946, *Ark. f. mat. astr. o. fysik*, **34B,** No. 1.
Maris, H. B., and Hulbert, E. O., 1929, *Phys. Rev.*, **33,** 412, 1046.
Martyn, D. F., 1951, *Nature, Lond.*, **167,** 92.
Maunder, E. W., 1904, *M.N.R.A.S.*, **64,** 205; 1905, *Ibid.*, **65,** 2, 538, 666; 1916, *Ibid.*, **76,** 63.
Maurain, C., 1927, *Ann Last. Physiques de Globe, Paris*, **5,** 87.
McLennan, J. C., 1928, *Proc. Roy. Soc.* A, **120,** 327.
McLennan, J. C., and Shrum, G. M., 1925, *Proc. Roy. Soc.* A, **108,** 501.
McNish, A. G., 1937, *Phys. Rev.*, **52,** 155.
Meinel, A. B., 1948, *Pub. Astron. Soc. Pacific*, **60,** 373; 1950 a, *Astrophys. J.*, **111,** 555; 1950 b, *Phys. Rev.*, **80,** 1096.
Milne, E. A., 1926, *M.N.R.A.S.*, **86,** 459.
Mitra, S. K., 1945, *Nature, Lond.*, **155,** 786.
Nagata, T., 1952, *Report of Ionosphere Res. in Japan*, **6,** 159.
Newton, H. W., 1943, *M.N.R.A.S.*, **103,** 246; 1944, *Ibid.*, **104,** 4.
Nicolet, M., 1948, *Gassiot Committee Conference* (London: Physical Society), p. 105.
Paschen, F., 1939, *Naturwiss.*, **18,** 752.
Photographic Atlas of Auroral Forms (Oslo 1930: International Geodetic and Geophysical Union).
Ratcliffe, J. A., and White, E. L. C., 1933, *Phil. Mag.*, **16,** 125.
Richardson, S. R., 1944, *Trans. Amer. Geophys. Union*, p. 558.

Rooney, W. J., 1934, *Terr. Mag. Atmos. Elec.*, **39,** 103.
Röstad, A., 1935, *Geofys. Publ.*, **10,** 1 (Oslo).
Schuster, A., 1911, *Proc. Roy. Soc.* A, **85,** 45.
Simpson, G. C., 1905, *Phil. Trans. Roy. Soc.* A, **205,** 92 ; 1933, *Quart. Journ. R. Meteorol. Soc., London*, **59,** 185.
Slipher, V. M., and Sommer, L. A., 1919, *Astrophys. J.*, **49,** 266 ; 1929, *Naturwiss.*, **17,** 802.
Sommer, L. A., 1930, *Naturwiss.*, **18,** 752.
Störmer, C., 1940, *Kristiania. Ske. Vid. Selsk*, **1,** 3 ; 1911, *Vid. Selsk, Skrift.*, **1,** (Mat. Nat. Kl. Oslo) ; 1911–12, *Arch. Sci. Phys. Geneva*, **32** ; 1929 a, *Nature, Lond.*, **123,** 82, 868 ; 1929 b, *Zeitschr. f. Geophys.*, **5,** 177 ; 1930, *Ibid.*, **6,** 463 ; 1937, *Nature, Lond.*, **139,** 584 ; 1938 a, *Ibid.*, **141,** 955 ; 1938 b, *Naturw.*, **26,** 633 ; 1942, *Geophys. Publ.*, **13** (Oslo).
Sugiura, M., Tazima, M., and Nagata, T., 1952, *Report of Ionosphere Res. in Japan*, **6,** 147.
Sverdrup, H. U., 1927, *Res. Depr. Terr. Mag., Carnegie Inst. Washington Publ.* 175, **6.**
Ta-You-Wu, 1943, *Proc. Ind. Acad. Sc.* (Bangalore), **18,** 40.
Vegard, L., 1916, *Ann. d. Phys.* (4), **50,** 853 ; 1921. *Phil. Mag.*, **42,** 47 ; 1928, *Handbuch der Experimentalphysik*, Band 25, Teil 1, p. 385 (Leipzig) ; 1930, *Se-Geophys.*, **6,** 42 ; 1932, *Geofys. Publ. Oslo*, **9,** No. 11 ; 1933, *Ibid.*, **10,** No. 4 ; 1936, *Nature, Lond.*, **138,** 930 ; 1938, *Ibid.*, **141,** 200 ; 1939, *Ibid.*, **144,** 1089 ; 1940, *Terr. Mag. Terr. Elect.*, **45,** 5.
Vegard, L., and Harang, L., 1933, *Geofys. Publ.*, **10,** No. 5.
Vegard, L., and Krogness, O., 1920, *Geofys. Publ.*, **1.**
Vegard, L., and Kvifte, G., 1945, *Geofys. Publ.* (Oslo), **16,** No. 7.
Vegard, L., and Tonsberg, E., 1937, *Geofys. Publ.*, **11,** No. 6.
Vestine, E. H., 1944, *Terr. Mag. Atmos. Elec.*, **49,** 77.
White, F. W. G., and Geddes, M., 1939, *Terr. Mag. Atmos. Elec.*, **44,** 367.

We beg to acknowledge the source of the following plates and text-figures :—

Photographs for plates 1–4 kindly supplied by Dr. J. Paton, The University, Edinburgh.
Plates 5 and 6 from Chapman, S., and Bartels, J., *Geomagnetism* (Plates 26, 28, 30, 31, 32 and 33).
Figure 1 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 462 (fig. 3).
Figure 2 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 460 (figs. 2 a, b).
Figure 3 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 465 (fig. 6).
Figure 4 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 469 (fig. 7).
Figure 5 from *Terrestrial Magnetism and Atmospheric Electricity*, 1939, **44,** p. 374 (fig. 1).
Figure 6 from Harang, L., *The Aurorae*, p. 9 (fig. 13) (or from Alaska).
Figure 7 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 474 (fig. 11).
Figures 8, 9 from Chapman S., and Bartels, J., *Geomagnetism*, p. 463 (figs. 4, 5).
Figure 10 from Harang, L., *The Aurorae*, p. 35 (fig. 46).
Figure 11 from Harang, L., *The Aurorae*, p. 10 (fig. 17).
Figure 12 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 475 (fig. 12).
Figure 13 from Harang, L., *The Aurorae*, p. 101 (fig. 107).
Figure 14 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 413 (fig. 9).
Figure 15 from Alfvén, H., *Kungl. Svenska, Vetenskap. Acad. Hand.*, Band 18, No. 3, p. 8 (fig. 2).
Figure 16 from Alfvén, H., *Kungl. Svenska, Vetenskap. Acad. Hand.*, fly sheet (fig. 3).
Figure 17 from Cowling, T. G., *Terrestrial Magnetism and Atmospheric Elect.*, 1942, **47,** (fig. 1).
Figure 18 from Alfvén, H., *Kungl. Svenska, Vetenskap. Acad. Hand.*, p. 18 (fig. 6).
Figure 19 from Alfvén, H., *Kungl. Svenska, Vetenskap. Acad. Hand.*, p. 21 (fig. 8).
Figure 20 from Alfvén, H., *Cosmic Electrodynamics*, p. 191 (figure on this page).
Figure 21 from Chapman .S., and Bartels, J., *Geomagnetism*, p. 856 (fig. 1).
Figure 22 from Chapman, S., and Bartels, J., *Geomagnetism*, p. 868 (fig. 5).

Infra-red Photo-conductors

By R. A. Smith
Telecommunications Research Establishment, Ministry of Supply, Malvern

CONTENTS

§ 1. Introduction

The progress in a field of research is seldom steady but more frequently takes place as a series of steps. A subject may be dormant or show only routine development for many years and then a sudden burst of activity takes place in which advances are made in a short time far greater than what has been achieved in the past decade. These sudden advances are

sometimes associated with the discovery of a new piece of fundamental information, but frequently they are associated with a development in experimental equipment which enables measurements to be made which had previously been impossible because of equipment limitations. Considerable advances in infra-red spectroscopy have been made during the last thirty years, especially after the importance of the infra-red spectrum in organic chemistry had come to be appreciated. Until quite recently, however, the detectors of infra-red radiation available to spectroscopists for the region of the spectrum between $1{\cdot}5\,\mu$ and $10\,\mu$, where most of the interesting spectra occur, had hardly shewn any advance for nearly half a century. For the near infra-red, say $0{\cdot}75\,\mu$ to $1{\cdot}5\,\mu$, several of the techniques widely used in the visible region of the spectrum (e.g., very sensitive photographic plates and photo-emissive detectors) had become available and a great increase in sensitivity had been achieved. Until the advent of the lead sulphide cell, however, only thermal methods of detection, with their limitations in speed of response, were available for wavelengths greater than $1{\cdot}5\,\mu$. The lead sulphide photo-conductive detector extended the use of photo-electric techniques to about $4\,\mu$, and recent developments with lead telluride and lead selenide detectors have pushed this limit to nearly $10\,\mu$. The use of these detectors has given a great impetus to high resolution infra-red spectroscopy. A review of recent work on the solar spectrum has been given by Goldberg (1950) and illustrates well the use of the new technique of using PbS detectors. Recent work on the fine structure of infra-red spectra by H. W. Thompson and his colleagues is typical of what now can be achieved with PbTe cells (Thompson and Williams 1952, Boyd and Thompson 1952). In addition to a reduction of the response time to a few microseconds, an increase in sensitivity of at least 100 times over the best thermal detectors is available with these cells (see § 7.2). It is this remarkable advance in sensitivity and speed that has brought about the recent developments in high resolution infra-red spectroscopy. A popular review of developments in the instrumentation of infra-red spectroscopy during the past half-century and in particular of the effect of the new photo-conductive detectors has been given by Strong (1951). Detailed reviews of both technique and types of spectra studied up to 1947 have been given by Williams (1948) and by Sutherland and Lee (1948). The present review will not be concerned with spectra but with the detectors and with the physics of the semi-conductors whose study led to their development.

As frequently happens the need for instrumental development leads to fundamental advances in physical knowledge, and again the advances in the latter lead to further instrumental development. The very large effort now concentrated on semi-conductor research, and in particular on germanium, arose through the development of the transistor, which was invented as a result of a fundamental study of the surface states on germanium. The development of the infra-red cells has been somewhat parallel. The need for infra-red detectors, with fast response, led to an

intensive study of PbS which has long been known to be photo-conductive in the infra-red. At first, the fundamental work had very little repercussion on cell development, which was mainly empirical. The study of PbS was extended to associated substances PbTe and PbSe, and their properties as semi-conductors became known. This has now led directly to the recent development of PbSe cells which extend the photo-electric infra-red spectrum to nearly 10 μ (see § 7.3). The present review will be concerned with the research carried out mainly since 1948 to study the properties of the infra-red photo-conductors as semi-conductors. Most of the work has been concerned with PbS, PbTe, and PbSe, but a considerable number of other substances has been studied, much less thoroughly. A review will then be given of infra-red cell development during this period.

A brief account of early work on infra-red photo-conductors and of the work carried out in Germany during the war has been given by Elliott (1947). This work led to the development of quite good PbS cells and to an appreciation that PbTe and PbSe were promising cell materials. It also shewed that these materials were semi-conductors which may exist as either p- or n-type but gave conflicting evidence on their electronic band structure. Early work in the U.S.A. and in particular the pioneering work of Case and Cashman has been discussed by Sutherland and Lee (1948). This work grew out of the development of thallium sulphide cells. These have a long-wave limit of about 1·1 μ and are not therefore effective in the spectral region 2 μ—10 μ with which we are concerned in this review. More recent American work, particularly that of Cashman on the development of PbS, PbSe, and PbTe cells has been briefly reviewed by Sutherland and Simpson (1952). A great deal of the work in this country on cell development has not been fully published but some account of the work at the Admiralty Research Laboratory, Cambridge University and the Telecommunications Research Establishment* (Ministry of Supply) has been given by Sutherland and Lee (1948) and in various publications referred to in their review. The early work on the physical properties of the PbS group of semi-conductors has been discussed by Smith (1951). It is therefore proposed, in the present review, to discuss mainly work carried out since 1947, both on semi-conductor research and on cell development. This will be divided as follows. First of all a brief account will be given of work on evaporated and chemically deposited layers and on other polycrystalline forms of PbS, PbTe and PbSe, followed by an account of similar work on other infra-red photo-conductors. Then an account will be given of the laboratory growth of single crystals of PbS, PbTe and PbSe and their use together with natural crystals of galena to study the electronic band structure of these substances and their fundamental properties as semi-conductors. The recent development of infra-red cells will then be discussed and the review will conclude with a theoretical discussion of the operation of these cells and of the ultimate limits of sensitivity of such detectors.

* Hereafter referred to as T.R.E.

§ 2. The PbS Group of Semi-conductors; Experiments with Thin Layers and other Polycrystalline Forms

The infra-red photo-conductors, such as PbS, form a very interesting group of semi-conductors and are worthy of fundamental study quite apart from their use as materials for the manufacture of photo-cells. They form an interesting contrast to the elementary semi-conductors such as germanium. They are similar in many respects and, in particular, are characterized by having a relatively low energy gap between the full and conduction bands. They can also exist as p-type and n-type, but their optical properties are somewhat different. Until recently these substances have only been available for study in polycrystalline form apart from fairly impure samples of natural galena. Because of their use in photo-cells in the form of evaporated or chemically deposited thin layers much of the early work was done on them in this form. One of the chief advantages of the thin layer is that various amounts of impurity in gaseous form may readily be added or extracted.

The main characteristics of any semi-conductor on which it is desired to obtain information are the width of the energy gap ΔE between the full and conduction bands and the nature, density and distribution of impurity levels between the bands, as a function of temperature. In addition it is of great interest to know the mobility of both electron and hole carriers and their effective mass. The measurements that may be made to obtain these with evaporated layers are very limited and it must be admitted that little progress was made till pure single crystals were available. Some of the features shewn by measurements on polycrystalline samples are however of interest and we shall briefly review these measurements.

The main type of measurement until quite recently was the variation of resistance and thermo-electric power with temperature. Provided it is assumed that the mobility of the carriers does not vary rapidly with temperature, the resistance R may be expected to vary as

$$R \simeq A \exp(\epsilon / 2kT). \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

For ' intrinsic ' conduction ϵ would correspond to the gap width ΔE between the full and conduction bands. For ' impurity ' conduction ϵ would correspond approximately to the depth of the impurity levels below the conduction band (n-type) or above the full band (p-type). Early measurements of the variation of resistance with temperature aimed at determining ΔE in this way. The sign of the thermo-electric power was usually taken as an indication of whether n-type or p-type conduction was predominant. There are a number of difficulties associated with this procedure. It is now well-known that the thermo-electric power is not a good index of the nature of the conduction and there is no convincing way of determining from the early measurements what is the significance of the value obtained for ϵ from an equation like (1). There is also a good deal of evidence to shew that much of the variation of resistance of evaporated layers with temperature arises from the barriers between the small

crystals of which the layer is composed. This will be discussed later. In order to avoid this difficulty experiments were also carried out with polycrystalline bulk samples and with crystals of natural galena. The crystals available to the early workers seem to have been rather impure and gave very variable results. When resistance–temperature measurements were reduced by plotting log R against $1/T$, in general the straight line expected from eqn. (1) was not obtained. A curve with positive slope corresponding to values of ϵ varying between 0·3 ev and zero and even becoming negative for some samples of PbS was obtained. The slope could be varied over a wide range of values by various treatments including baking in oxygen or in vacuo. This early work, particularly that of Bauer (1940), Eisenmann (1940) and Hintenberger (1942), established that the group of substances PbS, PbTe, PbSe are semi-conductors, can exist either as p-type or as n-type and indeed can be changed from the one to the other by suitable treatment. They can also behave as semi-metals. No convincing evidence for the value of ΔE was, however, obtained although it was thought to be about 0·3 ev for PbS (see § 3.2).

By measuring also the Hall constant at room temperature Hintenberger (1942) obtained values for the carrier mobility. These varied between 50 and 1 cm/sec per volt/cm. Such values are now known to be 20–1000 times lower than the mobilities found for single crystals of these substances (see § 3.2). This early work has been discussed by Chasmar and Putley (1951) who suggest that a large part of the variation of resistance with temperature in polycrystalline samples, and especially in layers, arises from the variation of contact resistance with temperature, the variation of resistance of individual micro-crystals being relatively small. It is found, and has been substantiated by later work on single crystals, that the more impure samples behave as semi-metals in the bulk and not as semi-conductors, i.e., they have a positive temperature coefficient of resistance.

2.1. *Structure of Thin Layers of PbS, etc.*

Recent work on evaporated layers has led to a better understanding of their nature. When viewed under a high-power microscope, layers as used in photo-conductive cells, having an average thickness of about 1 μ, appear as a series of ' islands ' made up of small crystals of sizes lying mainly between about 1 μ and 0·1 μ. This has been confirmed by Wilman (1948) using electron diffraction methods. The crystals making up the layer have the lattice parameters of the pure materials PbS, etc., within the accuracy of measurement. An interesting observation, which may be of significance in the theory of photo-conductivity of such layers, is that in layers treated with oxygen traces of the compound PbO . $PbSO_4$ appear. Measurements made at Purdue University confirm this and indicate that the PbO . $PbSO_4$ is confined to the surface of the crystals, since it shows up with electron diffraction but not with x-rays.*

* Private communication from Dr. K. Lark-Horowitz.

2.2. *Resistance of Thin Layers of PbS, etc.*

Having established the polycrystalline nature of such layers it is desirable to find the part played by contact resistance in determining the variation of electrical resistance with temperature. In order to do this, measurements have been made, on various types of layer, at frequencies up to 60 Mc/s by Chasmar (1948) who has shown that as the frequency is increased above 100 kc/s the resistance of the layer falls rapidly reaching a more or less steady value at about 60 Mc/s. Similar measurements have been made on PbS layers by Kolomiets (1948). These measurements were interpreted as shewing that at high frequencies the contact resistance is shorted out owing to the effects of barrier capacity and the resistance left is that of the bulk material of the crystallites making up the layer. If this is so, it would appear that less than 1% of the d.c. resistance is due to the latter. Moreover, the temperature coefficient of resistance decreases in magnitude as the frequency is raised and may even change sign and become positive when the d.c. value is negative. This would appear to shew that inter-crystalline barriers account for a large part of the variation of resistance of layers with temperature. It follows that the use of eqn. (1) together with log R/T^{-1} plots from measurements made on such layers are unlikely to give any useful information on the value of ΔE or on the distribution of impurity levels in the bulk material. This interpretation has, however, been criticized by Rittner and Grace (1952) who have shewn that a considerable decrease in the resistance of such layers at high frequencies is to be expected from the distributed capacitance of the film alone. Recent measurements at frequencies up to 200 Mc s by Humphrey, Lummis, and Scanlon (1952) have shewn that both distributive and intercrystalline capacity are required to explain the variation of resistance. There is, however, other evidence for the presence of barriers. For example, Simpson (1947) has shewn that if an evaporated layer of PbSe is compressed or extended 0·01% a reversible decrease or increase in resistance of about 5% is obtained. By assuming that the intercrystalline spacing is a linear function of the temperature and that electron conduction across the gap is by ' tunnel ' effect, he has obtained a theoretical expression for the variation of resistance of a layer with temperature which agrees fairly well with his observations. A similar model, in which conduction is by thermal excitation over the inter-crystalline barriers has been discussed by Chasmar and Putley (1951). By this means they have been able to account for the variation of resistance with temperature of a large number of evaporated layers. That a large part of the resistance, at least at room-temperature and below, in evaporated films is at inter-crystalline contacts can now hardly be doubted, as there is a great deal of other confirmatory evidence. For example, Starkiewicz, Sosnowski, and Simpson (1946) have observed marked rectifying effects when evaporated layers are ' polarized ' by passing current through them at temperatures of the order of 250°C. Marked photo-voltaic effects which vary in a random fashion across the layer are

also observed when it is examined by means of a very small spot of light (Sosnowski, Soole, and Starkiewicz, 1947). There is also some strong evidence obtained from a study of the photo-conductivity of such layers as we shall see later. The exact nature of the contacts is still somewhat in doubt and will be discussed in § 6.1.

When a layer of pure PbS about 1 μ thick is evaporated in a high vacuum a relatively low resistance film is obtained, corresponding to a bulk resistivity of about 0·01 ohm cm at room temperature. When a trace of oxygen is admitted the resistance rises extremely rapidly by as much as 1000 times. This may be due to the formation of high resistance inter-crystalline barriers. Further baking of the film in oxygen, in general, leads to a further increase of resistance and then a decrease under some conditions. This process has been studied by Sosnowski, Starkiewicz and Simpson (1947) and by Schwarz (1948). The former have shewn that under certain conditions a layer which starts as n-type, as indicated by thermo-electric power, is gradually changed to p-type, the resistance passing through a maximum as the thermo-electric power passes through zero. This does not always take place and conversion of a layer to an oxide form may be obtained with continuously increasing resistance. The effect of oxygen on the resistance of PbS films has been studied by Ehrenberg and Hirsch (1951) using an ingenious apparatus in which the temperature is varied so rapidly that changes in the structure of the layers are unlikely during the measurement. In the interpretation of their measurements they have entirely ignored the effect of inter-crystalline barriers. Quite recently it has been shewn by Putley (1952 b) by making a direct comparison between single-crystals and sintered polycrystalline samples of PbS, that, except at quite high temperatures (above 600°K), resistance–temperature measurements give quite different results. This work would appear to shew conclusively that in such measurements, except at high temperatures, the inter-crystalline barriers play a prominent part in determining the variation of resistance with temperature.

2.3. *Hall Constant Measurements on Evaporated Layers of PbS, etc.*

Since the early Hall constant measurements of Hintenberger (1942), using electrometer techniques, which shewed that the mobilities of electrons and positive holes in evaporated layers of PbS, etc., were of the order of 1–50 cm/sec per volt/cm, very little has been done until recently. The method used by Hintenberger is difficult to apply to high resistance layers and it is hard to avoid the effects of other thermo-magnetic effects. An alternating current technique which overcomes some of these difficulties has recently been used by Lothrop (1949) and Halvorsen (1951) at Northwestern University, U.S.A. Another method has been developed by Russell and Wahlig (1950) and has also been used at T.R.E. This uses different frequencies for the magnetic field and current passing through the specimen, and the Hall voltage is observed at the sum or difference frequency. Measurements have shewn that films can be converted from n-type to p-type and vice-versa with various

treatments in oxygen and in vacuo and may change character over different temperature ranges. Mobilities of the same order as measured by Hinterberger are also found. The measurements are difficult to interpret due to the effect of inter-crystalline barriers. Putley (1952 b) has, however, shewn recently that poly-crystalline sintered samples give similar values for the Hall constant to those obtained with single crystals of comparable purity. Although it has not yet been definitely established, there is reason to hope that Hall constant measurements on evaporated layers may also give valuable information regarding the bulk properties of the crystals of which the layer consists, whereas, as we have seen, resistance measurements are unlikely to do so.

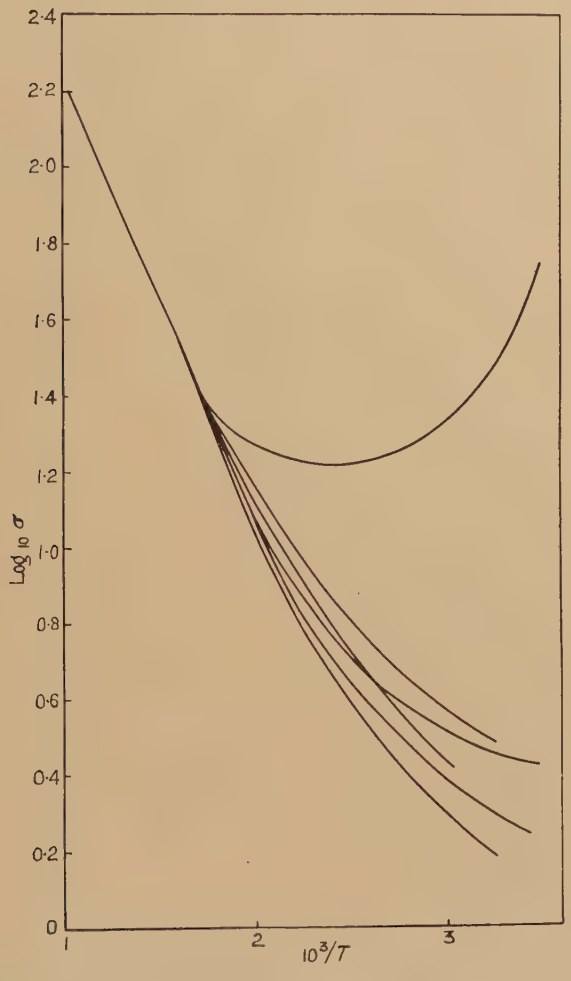
2.4. *Resistance and Hall-constant Measurements on Poly-crystalline Bulk Samples of PbS, etc.*

Since the early measurements of Bauer (1940), Eisenmann (1940), and Hintenberger (1942), using polycrystalline ingots cooled from a melt, very little progress with measurements on such samples has been made until fairly recently. The early measurements shewed that such samples had a very much lower resistivity at room temperature than is generally obtained with evaporated layers—of the order of 0·1–0·001 ohm cm. These samples shewed in some cases negative temperature coefficients of resistance and could exist as either n-type or p-type as shewn by Hall constant measurements. Frequently positive temperature coefficients of resistance were found and at one time it was wondered whether indeed these substances were semi-conductors in the bulk, the negative temperature coefficients sometimes observed being possibly due to inter-crystalline barriers. It is now known that such behaviour is characteristic of impure samples and that pure samples do indeed behave as semi-conductors. No clear indication of intrinsic conductivity was observed from the slopes of the log R/T^{-1} curves which varied greatly from one sample to another. This is now known to be due to the fact that insufficiently pure samples were available. The first clear evidence of intrinsic conductivity in this group of substances was obtained by Chasmar and Putley (1951) for PbTe. They measured the variation of resistance and Hall constant for a number of poly-crystalline ingots over a temperature range extending from 1000°K to 300°K. The ingots were of fairly high purity and resulted from unsuccessful attempts to grow single crystals. Both n-type and p-type samples at room temperature were used. At temperatures below 600°K the resistance and Hall constant values varied greatly from sample to sample but above this temperature very nearly the same value was found for all samples. Curves giving the variation of conductivity σ and Hall constant $\mathscr{R}$ are shewn in figs. 1 and 2. The high temperature parts of the curves were interpreted as indicating intrinsic conductivity. In this range both the resistance and Hall constant are inversely proportional to n_e the number of conduction electrons, which is practically equal to the number of ' holes '. Also (Mott and Gurney 1940)

$$n_e \simeq \text{constant } T^{3/2} \exp(-\Delta E/2kT) \quad . \quad . \quad . \quad . \quad . \quad (2)$$

so that a logarithmic plot will give ΔE. It will be seen from figs. 1, 2, that the slopes obtained from such plots are approximately constant in the high temperature range and that of the Hall constant curve leads to a value of $\Delta E = 0{\cdot}6$ ev. That this value is approximately correct has since been verified using single crystals (see § 3.2). Moreover Putley (1952 b) has

Fig. 1

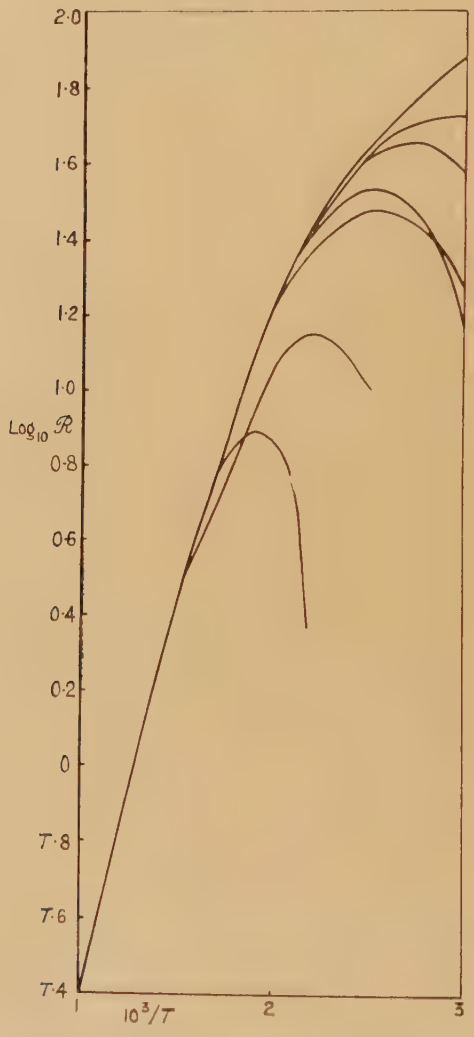


Conductivity of PbTe (Chasmar and Putley 1951).
(σ in ohm^{-1} cm^{-1}, T in °K.)

recently shown that in the intrinsic range sintered samples of PbS give approximately the same value for ΔE as single crystals. Resistance values at lower temperatures are, however, quite different for the two types of sample. It thus appears that resistance measurements using polycrystalline samples may give very misleading values. Fortunately it turns out that the Hall constant measurements for the two types of sample give

almost identical results and this type of measurement may thus be likely to give a more reliable value of ΔE. It may be noted that the value of ΔE obtained from fig. 1 is slightly different from that obtained from fig. 2. This is thought to be due to the variation of m bility with temperature which is included in the variation of resistance as a first

Fig. 2



Hall constant for PbTe (Chasmar and Putley 1951).
($\mathscr{R}$ in cm^3/coulomb, T in °K.)

order effect. This is another reason for expecting a more reliable value of ΔE from the Hall constant measurements. The value found for PbS is about 1·2 ev. (For more detailed measurements using single crystals see

§ 3.2. We may also note here that for PbSe $\Delta E = 0{\cdot}5$ ev.) The Hall constant for the intrinsic range turns out to be negative. This shews that conduction is predominantly by electrons so that the mobility of electrons is greater than that of ' holes ' since, in this range, their numbers are nearly equal. The values for the mobility deduced from these experiments are of the order of 1000 cm/sec per volt/cm and are 100–1000 times greater than those obtained for evaporated layers.

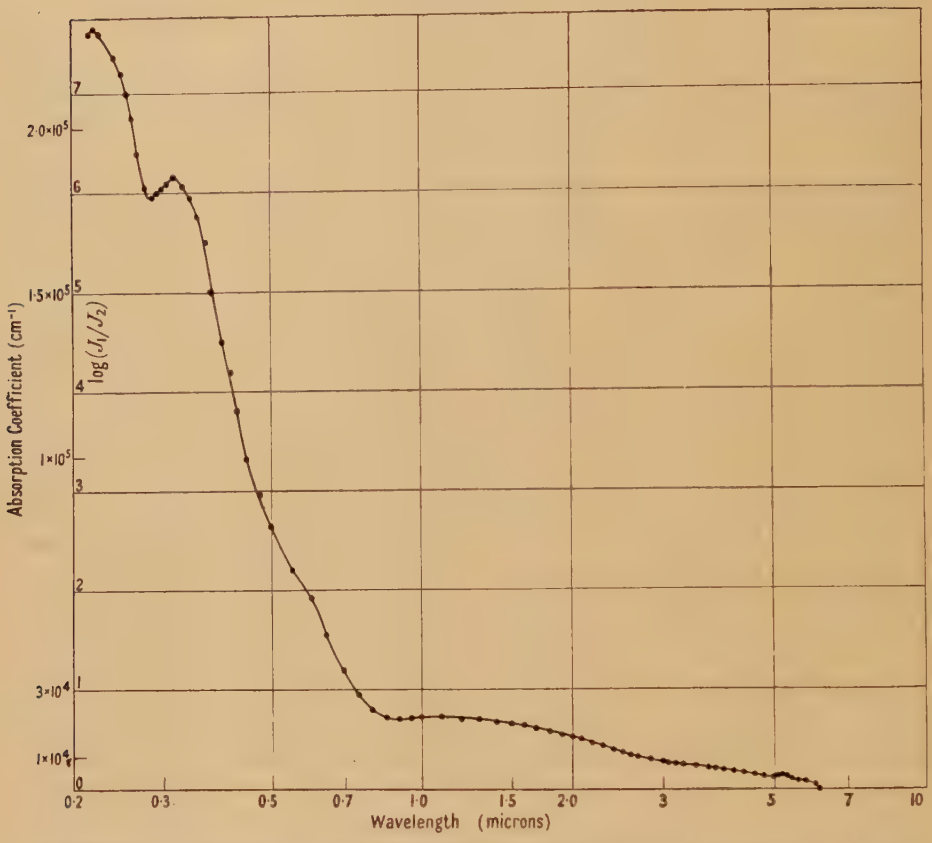
2.5. *Infra-red Absorption of Evaporated Layers of PbS, etc.*

Another method of obtaining information on the value of the forbidden energy gap ΔE, and possibly on the distribution of impurity levels, is to observe the absorption spectrum of the semi-conductor in question. This method, while giving useful confirmation of values obtained by the electrical measurements discussed above, cannot by itself give an unambiguous result owing to the difficulty in deciding exactly which electronic transition is associated with a particular absorption band or edge. In common with most semiconductors the absorption of the PbS group is very high (10^5–10^6 cm^{-1}) in the visible and near infra-red. Consequently most of the measurements in this region of the spectrum have been obtained with evaporated or chemically deposited layers of the order of $0{\cdot}1$–$10\,\mu$ thick. The most extensive measurements have been made by Gibson (1950) who has studied the absorption of layers of PbS, PbSe, and PbTe produced in various different ways. He has also studied the variation of absorption with temperature. The absorption spectrum at room temperature of a typical layer of PbS is shewn in fig. 3. It will be seen that as the wavelength is decreased beyond about $1{\cdot}0\,\mu$ the absorption increases rapidly and reaches a value of the order of 10^6 cm^{-1} in the visible and near ultra-violet. This absorption ' edge ' is interpreted as being accociated with transitions of electrons from the full to the conduction band. If this is so, and the high value of the absorption coefficient would support this, the value of ΔE must correspond to a quantum energy of approximately 1·2 ev. It is always difficult to obtain an exact value of ΔE from an absorption ' edge ' of this kind as it is rarely sharp. The value is, however, in good agreement with Putley's value 1·2 ev obtained from Hall constant and conductivity measurements (see § 2.4).

To the long-wave side of the absorption edge there is a long ' tail ' extending to $6\,\mu$, and shewing a relatively high absorption coefficient of the order of 10^4 cm^{-1}. Beyond $6\,\mu$ the absorption is a good deal less but these experiments do not give a value. It must be admitted that, so far, no really satisfactory explanation has been given for this absorption which, as we shall see later, cannot be associated with the transitions leading to photo-conductivity, or indeed with any other known transitions. Moreover this absorption is found to be largely absent in single crystals (see § 3.3). It is probably due to scattering by the micro-crystals of which the layer consists.

When the temperature is decreased it is found that the absorption edge near $1\,\mu$ moves towards the *long-wave* end of the spectrum, the rate of shift corresponding to an energy change of about 6×10^{-4} ev/°K. At the same time the limit of the ' tail ' moves towards the short-wave end and at 77°K comes to just over $3\,\mu$ (see fig. 4).

Fig. 3



The absorption spectrum of a chemically deposited layer of PbS (Gibson 1950).

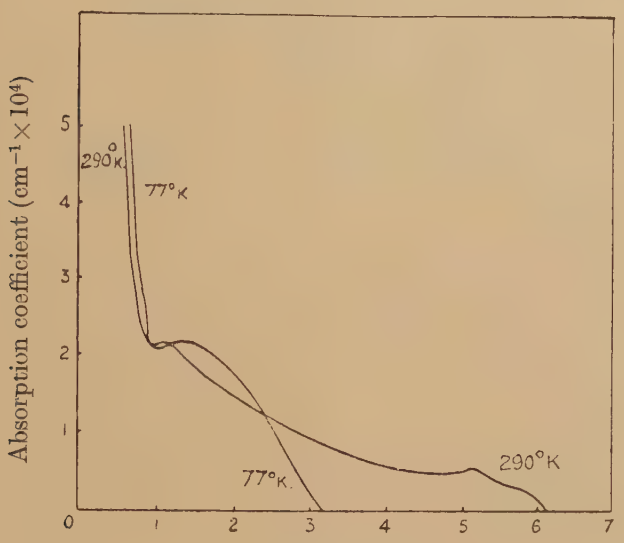
Very similar absorption curves and variation with temperature were found for layers of PbSe and PbTe. In this case the beginning of the absorption edge is not so well marked and comes about $1{\cdot}5$–$2\,\mu$ corresponding to quantum energies of 0·8–0·6 ev, slightly higher than the values obtained by Putley (see § 3.2). They are, however, sufficiently near to make it likely that the absorption associated with the ' edges ' corresponds to transitions from the full to the conduction band. In this case the long-wave ' tails ' extend to beyond $12\,\mu$ and absorption coefficients in the ' tails ' as high as 10^5 cm^{-1} are obtained with the purest samples used. When oxygen was baked into the layers the absorption in the ' tails ' decreased, but only to about half its highest value. The absorption to the

short-wave side of the ' edge ' was, however, unaffected by such treatment. We may remark here that no change in absorption in such layers was observed at a wavelength corresponding to the long-wave limit of photo-conductivity (see § 2.6).

2.6. *Photo-conductivity in Thin Layers of PbS, etc.*

In view of the practical importance of the infra-red photo-conductivity of the PbS group of substances more observations have been made of this property than of any of the other properties of these semi-conductors. It must, however, be admitted that, as a result, very little has been found out about the more fundamental properties of these substances. The

Fig. 4



Variation of absorption of PbS layer with temperature (Gibson 1950).

early work in Germany and in the U.S.A. shewed that these substances appeared to be unique in showing very marked photo-conductivity at much longer wavelengths than any other known photo-conductors. PbS, for example, was known to be photo-conductive to about 3·5 μ at room temperature and to about 4·5 μ at liquid air temperature. PbTe and PbSe were thought to be photo-conductive to about 6 μ at liquid air temperature. Within recent years much more detailed studies have been made of the spectral variation of photo-conductivity over a wide range of temperatures. The results for PbS, PbTe, PbSe will be described in the following sections.

2.6.1. *Photo-conductivity of PbS.*

The most extensive recent measurements of the photo-conductive response of PbS have been made by Moss (1947, 1949 b). Figure 5 shews the form of response taken with an equal energy spectrum. The chief

characteristic is that at wavelengths greater than a certain value the response falls off rapidly but there is no point at which it can be said to be zero. In practice, the photo-conductive response becomes unobservable at a certain wavelength but this depends on the ultimate sensitivity of the experimental arrangement, the limit being set by fluctuation noise which may be reduced by narrowing the bandwidth. If the variation of response

Fig. 5

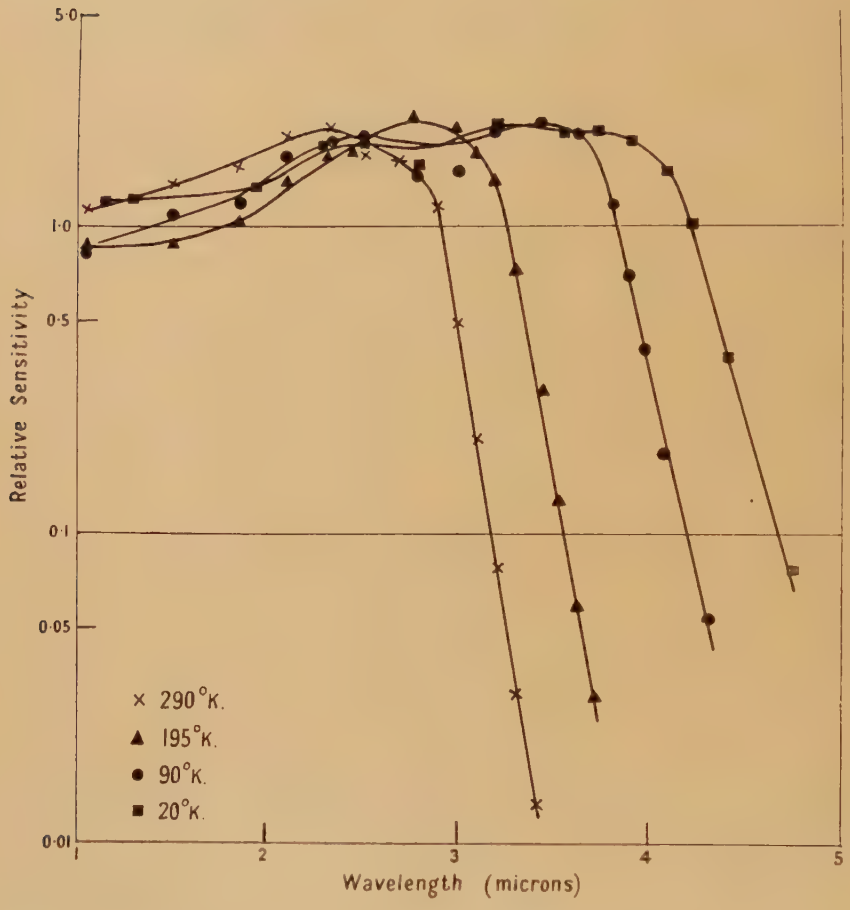


Photo-conductive response of PbS (Moss 1949 b).

with wavelength is plotted on a logarithmic scale (fig. 5), this rapidly falling part of the curve is very nearly linear. This is a very characteristic feature of such curves, and if not observed, generally means that the sensitivity is insufficient to sh w up this part of the curve. It is desirable to be able to fix a definite ' long-wave limit ' for the photo-conductivity. Various methods have been suggested, but the one now fairly generally

adopted is that suggested by Moss (1952 b), namely the wavelength at which the response has fallen to half the value of the maximum occurring just before the rapid fall. This is denoted by $\lambda_{1/2}$. According to Moss this value has a theoretical significance for comparison with thermal and other energies associated with the photo-conductive transition.

From fig. 5, another very characteristic feature will be evident. As the temperature is decreased the 'long-wave limit' moves to longer wavelengths. This is contrary to what is observed for most other photo-conductors. The shift is found to correspond to an energy change of $4{\cdot}7 \times 10^{-4}$ ev/°K over a wide range of temperature.

At one time it was thought that the value of $\lambda_{1/2}$ varied very considerably with state of purity of the photo-conductive layer (Smith 1951). Considerable experience with sensitive layers as used in detectors has now led to the belief that this variation is fairly small. Some of the marked changes observed previously are now thought to arise simply from the fact that the photo-conductivity near the $\lambda_{1/2}$ value had fallen below observation level whereas that at much shorter wavelengths was still high. This is observed in layers which have either been inadequately sensitized with oxygen or other treatment or which have been strongly oxydized. For example photo-sensitivity curves published recently for layers of widely different character including evaporated layers (Clark and Cashman 1952) and chemically deposited layers (Eastman Kodak Co. 1952, Milner and Watts 1952) shew responses very similar to those obtained by Moss. Values for $\lambda_{1/2}$ now commonly found are $2{\cdot}9\,\mu$ for room temperature or $3{\cdot}75\,\mu$ for liquid air temperature.

An interesting feature of the spectral response curves obtained with modern sensitive photo-conductive cells is that at wavelengths shorter than that corresponding to maximum response the response for an equal energy spectrum falls off slowly and regularly. If instead the response per incident quantum is plotted against wavelength this part of the curve shews an almost constant value from wavelengths of the order of $1\,\mu$ right up to the point where the response begins to fall rapidly. There is also some evidence (see § 7.4) that the quantum efficiency for this region is not far from unity.

It is now well established that layers showing the above behaviour, and in particular the high photoconductive sensitivity out to the limits quoted, can only be obtained after suitable treatment such as baking in a low pressure of oxygen. Very little information has been published on the details of the processes used and these vary considerably with individuals making photo-conductive layers in the laboratory and with commercial organizations making cells in numbers.

Two main methods are now used in the preparation of sensitive layers, evaporation in a low pressure of oxygen or in vacuo with subsequent baking in a low pressure of oxygen (Cashman 1946, Sosnowski, Starkiewicz, and Simpson 1947) and chemical deposition (Kicinski 1948) in the presence of an oxidizing agent. An electric discharge has also been used during evaporation (Schwarz 1948).

When a layer of PbS of the highest purity available is evaporated in a vacuum better than 10^{-6} mm Hg its photo-sensitivity is small or nil. This indicates that the very marked photo-sensitivity is due to a secondary process. The fact that the long-wave limit is characteristic of the basic material appears to shew, nevertheless, that a primary process is also involved. This will be discussed in § 6.1. It appears that the oxygen or other treatment serves to show up the basic phenomenon by greatly enhancing its effect.

That a secondary process is taking place is also indicated by the fact that the speed of the photo-conductive response to a sudden change of radiation intensity varies widely with treatment. With evaporated layers, time constants at room temperature may vary with treatment from one or two microseconds to several hundred microseconds. For chemically deposited layers the time constants are usually longer, ranging from a few hundred microseconds to greater than a millisecond (see § 7). An extensive series of experiments on the variation of time constant with various parameters for PbS layers has been carried out by Gibson (1950). He found that the decay in photo-conductivity was generally made up of two exponentials, corresponding to a rapid initial fall followed by a slower decay. The initial time constant is the important one in practice and Gibson found this to vary in a very marked manner with temperature. Below a certain temperature, characteristic of each layer, but generally about 200°K, the initial time-constant varies very little. Above this temperature, however, it decreases rapidly with increasing temperature in an approximately exponential fashion. The significance of this in understanding the theory of photo-conductivity of layers will be discussed in § 6.1. It is also found that the time-constant decreases rapidly when the current through the layer exceeds a certain value and also at high intensities of illumination.

A number of other curious phenomena associated with photo-conductivity have been observed by Gibson (1949) and by Chasmar and Gibson (1951). It was found, for example, that if a layer of PbS has been rendered strongly p-type by heating in sulphur vapour, or in oxygen, to such an extent that normal infra-red photo-conductivity has been practically destroyed it can be rendered photo-conductive again in the infra-red if illuminated for some time at 90°K by visible light. During illumination the resistance steadily decreases. The light required to produce this effect must have a wavelength shorter than the absorption edge of 1·1 μ which is believed to correspond to excitation of electrons from the full to the conduction band. The photo-sensitivity decays slowly if the specimen is kept at 90 K but disappears quickly if it is warmed up to room temperature.

2.6.2. *Photo-conductivity of PbTe*

Evaporated layers of PbTe may be made to shew marked photo-conductivity by oxygen treatment just as for PbS. The spectral response for such evaporated layers has also been measured by Moss (1948, 1949 b).

The curves for variation of photo-sensitivity with wavelength are very similar to those obtained with PbS except that the $\lambda_{1/2}$ value for a given temperature is shifted to longer wavelengths (see fig. 6). A rather different variation with wavelength has been found by Simpson, Sutherland, and Blackwell (1948) and by Simpson and Sutherland (1951) who endeavoured to exclude oxygen from their layers. This, however, is not easy without the use of very advanced high-vacuum techniques. A good deal of experience at T.R.E. with sensitive photo-conductive PbTe cells has shewn that curves similar to those obtained by Moss are always found for well sensitized layers, which have been treated with oxygen. Curves recently published by Clark and Cashman (1952) are also of this form. Like those for PbS they shew a sharp linear drop when plotted on a logarithmic scale at wavelengths greater than a certain

Fig. 6

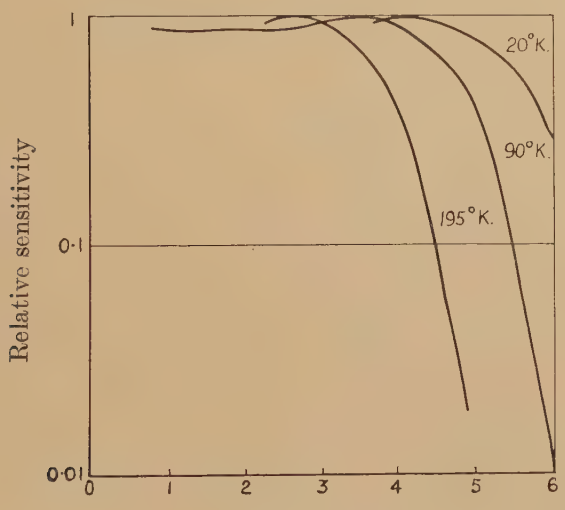


Photo-conductive response of PbTe (Moss 1949 b).

value. At shorter wavelengths a slow fall is shewn, when the response is plotted for an equal energy spectrum, but again the response per quantum is almost constant till the rapid decrease occurs. The value of $\lambda_{1/2}$ for PbTe at 90°K is about 4·75 μ and the spectral shift to longer wavelengths as the temperature is lowered corresponds to an energy change of approximately $5{\cdot}0 \times 10^{-4}$ ev/°K. PbTe differs from PbS in that very little sensitivity is observed at room temperature, it being necessary to cool the layer with solid CO_2 before any appreciable response is obtained. Only evaporated layers of PbTe have been studied, no chemical process for forming such layers having been described.

The time-constants associated with photo-conductivity in layers of PbTe have not been so extensively studied as for PbS. The behaviour is

however, very similar. Measurements by Scanlon, Petritz, and Lummis (1952) have shewn the existence of double time-constants as found for PbS by Gibson (1951).

2.6.3. *Photo-conductivity of PbSe*

Both the early work (see Sutherland and Lee 1948) and several measurements within recent years have shewn that apparently the long-wave limit of photo-conductivity of PbSe is the same as that for PbTe, or occurs at a slightly shorter wavelength. This is what we should expect since Se lies between S and Te in the periodic table. These measurements have been carried out both with evaporated layers (Blackwell, Simpson, and Sutherland 1947, Moss and Chasmar 1948, Starkiewicz 1948, Moss 1949 b) and with chemically deposited layers (Milner and Watts 1949). The

Fig. 7

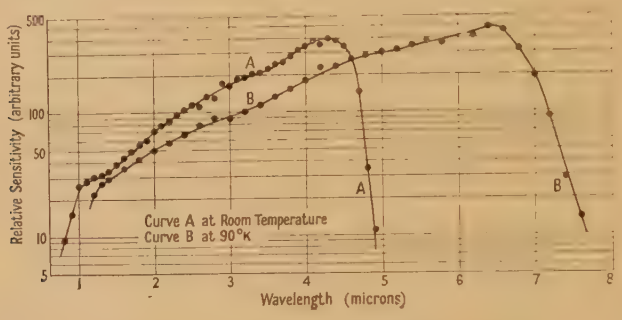


Photo-conductive response of PbSe (Gibson 1952).

A Room temperature.
B 90°K.

long-wave limit has, however, recently been shewn by Gibson, Lawson and Moss (1951) to be at a considerably longer wavelength than for PbTe (fig. 7). This surprising result was first obtained by use of single crystals (see §§ 3.3, 3.4) but was later verified for evaporated layers. The $\lambda_{1/2}$ value for PbSe is $4{\cdot}7\,\mu$ for room temperature, $7{\cdot}1\,\mu$ at 90 K and $8{\cdot}2\,\mu$ at 20°K. This corresponds to an energy shift of about 4×10^{-4} ev/ K. One can only assume that the layers on which previous measurements had been made were not sensitized sufficiently to shew the long-wave photo-conductivity, which work on single crystals has shewn to be characteristic of the material. It may be noted that as for PbS, but not PbTe, appreciable photo-conductivity is observed at room temperature.

The form of spectral response found by various workers varies a good deal and shews peaks in different parts of the spectrum. For well sensitized layers such as used for photo-conductive cells at T.R.E. the same form of curve as for PbS and PbTe is found (see § 7.3). This shews a nearly constant response per quantum till a certain wavelength is reached and falls off exponentially at longer wavelengths. The time constants for PbSe are similar to those for PbTe but tend to be slightly smaller.

2.7. *Measurements of Thermoelectric Power*

It was established by most of the early workers (e.g., Hintenberger 1942) that the thermo-electric power of all the substances of the PbS group could be either positive or negative and that the sign could be changed by various heat treatments. For example if a tellurium-rich sample of PbTe is heated in vacuo the sign of the thermo-electric power will change from positive to negative. The reverse process takes place if it is then heated in oxygen.

Observations on PbS evaporated layers of Sosnowski, Starkiewicz and Simpson (1947) have shewn that baking a lead-rich film of PbS in oxygen causes the resistance to rise and then fall, the maximum being reached as the thermo-electric power changes sign. Similar results have also been obtained by Wyrich and Levinstein (1950) for PbTe layers. Experience in making photo-conductive cells at T.R.E. has shewn, however, that this behaviour is not found under all conditions and that the observed thermo-electric power is not always a reliable index of the sign of the predominant carriers.

§ 3. The PbS Group of Semi-conductors: Experiments with Single Crystals

Some of the early work on PbS was carried out with natural crystals of galena. It is not clear whether or not these samples were in the form of single crystals. They were, in general, too impure to sh w intrinsic conductivity. The first convincing evidence for this was obtained for this group of substances, as we have seen (§ 2.4) by Putley and Chasmar (1951) using polycrystalline but fairly pure samples of PbTe. Natural crystals of reasonable dimensions and purity of PbTe and PbSe are not available although these substances do exist naturally, in small quantities, in crystalline form as altaite and clausthalite. More recent experiments with natural crystals of PbS by Dunaev and Moslakovitz (1947) failed to give convincing evidence of intrinsic conductivity. This was found, however, by Putley and Arthur (1951) using samples of galena from Sardinia of exceptionally high purity.

3.1. *Growth of Single Crystals of PbS, etc*

Since natural crystals of PbS only were available for experiment, and the purity of these is not under control, an attempt was made at T.R.E. to grow crystals of all three of the substances of this group in the laboratory. Since PbS crystals exist abundantly in nature as galena it was thought that of the three substance of this group PbS would be easiest to grow. Attempts to grow crystals of appreciable size, first by condensation from the vapour phase, and later by the Bridgman-Stockbarger method of slowly dropping a melt through a freezing plane, were unsuccessful. On turning to PbTe, however, Lawson (1951, 1952) was able to grow large single crystals using the latter method. It was found, at first, that in spite of variations of the ratio of Pb to Te in the original mix, these crystals always came out p-type. It was then discovered that this was due to the

presence of oxygen. After careful elimination of oxygen by reduction with hydrogen prior to melting, Lawson was able to grow good crystals of either n-type or p-type. On applying this improved technique to PbS he was then able to grow good single crystals of this substance as well, and later of PbSe. Details of the apparatus and the various precautions necessary will be found in Lawson's papers. This successful growth of single crystals has led, as we shall see, to some very interesting new observations on their optical absorption and photo-conductivity. Crystals of PbTe have also been recently grown by Clark and Cashman (1952).

3.2. *Conductivity and Hall Constant Measurements on Single Crystals of PbS, etc.*

Measurements of the variation of conductivity and Hall constant with temperature on some natural crystals of galena by Putley and Arthur (1951) gave a series of curves very similar in form to those obtained by Chasmar and Putley (1951) with pure polycrystalline samples of PbTe (figs. 1, 2). The slope of the Hall constant curves at high temperatures, in this case, gave a value for the energy gap of PbS of 1·17 ev (as compared with 0·6 ev for PbTe). By using single crystals grown by Lawson (1951, 1952), Putley (1952 a, b, c) in a series of important experiments, has verified these results generally and has obtained accurate values for the energy gap ΔE for PbS, PbSe, PbTe at temperatures above 500°c.* In the intrinsic range he finds that the value of the conductivity is similar for bulk polycrystalline samples and for single crystals of comparable purity. This is not so, however, for lower temperatures, the values being consistently higher for the single crystals. This is undoubtedly due to the effect of inter-crystalline barriers. The Hall constant measurements give similar results for both types of sample as has also been verified by the direct comparison between single crystals and sintered samples of PbS as discussed in § 2.4.

For the single crystals the general behaviour of the conductivity is as follows. For relatively pure samples (having about 10^{16} effective donors or acceptors per c.c.) the conductivity increases rapidly with increasing temperature above 500°K, is fairly constant around room temperature and generally increases as the temperature is further decreased. From Hall constant measurements the number of carriers increases rapidly with increasing temperature above 500°K, and is fairly constant below room temperature. For p-type samples a reversal in the sign of the Hall constant usually takes place between these temperatures. The temperature above which the conductivity becomes intrinsic depends on the number of impurities, and for impure samples (with say $>10^{18}$ impurities per c.c.) the conductivity may decrease as the temperature is raised right up to 900°c, at which temperature marked and irreversible changes in characteristics take place. In the intrinsic range we have theoretically

$$\mathscr{R}=A\,T^{-3/2}\exp\left(\Delta E/2kT\right)$$

* See footnote p. 341.

and if log ($\mathscr{R}T^{3/2}$) is plotted against $1/T$ a fairly good but not exactly straight line is usually obtained, from which a value of ΔE can be deduced. In this way the following values of ΔE have been obtained.*

PbS $\Delta E = 1{\cdot}17$ ev,
PbSe $\Delta E = 0{\cdot}5$ ev,
PbTe $\Delta E = 0{\cdot}63$ ev.

These may be regarded as the best available values of ΔE for the intrinsic range above 500°C.* That for PbSe is probably less accurate than the others owing to the difficulty of obtaining sufficiently pure crystals of this substance, but it is certainly considerably less than that for PbTe—a most surprising result, which was, however, anticipated by measurements on absorption (see § 3.3). The above values are in general agreement with the rapid decrease in absorption found for evaporated layers by Gibson (1950) (see § 2.5).

From these measurements values of the mobility have also been deduced. It is generally found that the mobilities for single crystals are several times greater than for polycrystalline bulk samples, again shewing the effect of barriers. For the single crystals remarkably high values of the mobility are obtained. The *highest* values observed at room temperature are as follows :—

		PbS	PbSe	PbTe
Max. observed mobility for 290°K in cm/sec per volt/cm	electrons	640	1400	2100
	holes	800	1400	840

The values quoted for electrons and holes apply, of course, to different crystals.

Since the value of the Hall constant is always negative in the intrinsic range it follows that the electron mobility is higher than the hole mobility. It is probably at least two times higher. From the above maximum values we should therefore expect considerably higher electron mobilities to be observed. The mobility does not appear to vary much with impurity content and from his analyses Putley concludes that there is no evidence for impurity scattering in these compounds. At lower temperatures a great increase in mobility is observed. For example values of about 10^4 cm/sec per volt/cm are observed at 77°C and 10^5 cm/sec per volt/cm at 20°K. Over a temperature range of about 700°K–100°K it is found that the mobility μ is well represented by an equation of the form

$$\mu = \mu_0 T^{-5/2}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

---

* *Note added in proof*:—The above values of ΔE, obtained from the slope of curves similar to those of fig. 2, may be too high. If the more complex expression for $\mathscr{R}$ which is got by considering mixed conduction is used, together with measured values of the mobility of holes and electrons obtained for various temperatures by use of fairly impure p- and n-type samples respectively, somewhat smaller values of ΔE are found. This does not, however, reduce the value of ΔE for the observed intrinsic range to the order of 2·4 ev for PbS. The author is indebted to Dr. E. H. Putley for this information.

So far, there is no theoretical basis for such a variation. We may note that these mobilities are much greater than those obtained for evaporated layers (0·5–50 cm/sec per volt/cm). Also the mobilities for layers generally decrease as the temperature is lowered.

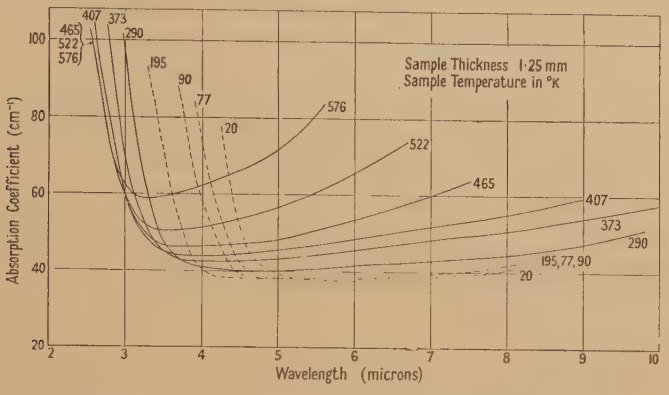
3.3. *Absorption Measurements with Single Crystals of PbS, PbSe, PbTe*

Until recently measurements on the absorption of bulk samples of all three substances of this group appeared to shew that they were quite opaque in the ultra-violet, visible and infra-red regions of the spectrum. Gibson's (1950) measurements on thin films appeared to confirm this, at least out to wavelengths of the order of 6 μ where absorption coefficients greater than 10^4 cm^{-1} were found. The first observations which shewed that the absorption in certain single crystals might be much less than had previously been supposed were made by Paul, Jones and Jones (1951) who shewed that in an exceptional crystal of natural galena, the absorption coefficient was as low as 5 cm^{-1} at wavelengths around 4 μ. At wavelengths shorter than 3 μ the absorption was observed to rise extremely rapidly. This absorption 'edge' corresponds very well with the long-wave limit of photo-conductivity. Although carefully looked for with thin films, no indication of its presence had been found (Gibson 1950). For films of the type studied by Gibson (1950) it is clearly masked by a much greater absorption ($\sim 10^4$ cm^{-1}) whose cause is at present uncertain and which is apparently absent in single crystals. Such an absorption edge was also found by Paul, Jones, and Jones (1951) for a T.R.E. synthetic crystal of PbSe. Following on this interesting discovery it was soon found (Gibson, Lawson, and Moss 1951) that all three substances of this group PbS, PbSe, PbTe in the form of single crystals grown by Lawson shew considerable transmission at wavelengths beyond an absorption edge whose position corresponds very well with the long-wave limit of photo-conductivity as observed with sensitized layers. We may note here that bulk photo-conductivity had not yet been observed. This correspondence has also been observed for natural galena specimens and a laboratory-grown PbTe crystal by Clark and Cashman (1952), who have also shewn that the temperature variation of the position of the absorption edge is the same as that of the long-wave limit of photo-conductivity.

A detailed investigation of this absorption and its variation with temperature and with impurity content of the specimens has been made by Gibson (1952 *c*) using T.R.E. laboratory-grown single crystals of PbS, PbSe, PbTe, and by Paul and Jones (1953) using galena crystals. These experiments confirm in detail the correspondence over a wide range of temperatures between the photo-conductive 'limit' as found with thin-layer photo-cells and also with point contact photo-cells made with single crystals (see § 3.4) and the absorption edge, for PbS and PbTe. The results found for PbSe were, however, at first very surprising. A value of about 5 μ at room temperature and 7 μ at liquid air temperature was found

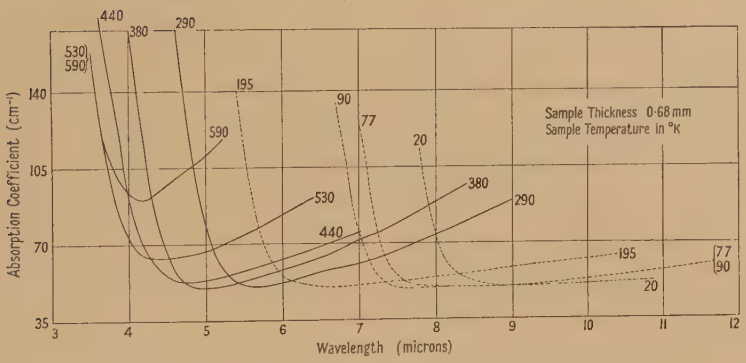
for the position of the absorption edge. From measurements on layers the corresponding photo-conductive 'limits' were thought to be about 3·5 μ and 5 μ. It was, soon shewn, however, using both single crystal photo-cells and evaporated layers that these values were in error and that the correspondence in the case of PbSe was also complete (Gibson, Lawson, and Moss 1951). That these wavelengths should exceed the corresponding ones for PbTe by a considerable amount is most surprising.

Fig. 8



Absorption of n-type PbS crystal (Gibson 1952).

Fig. 9



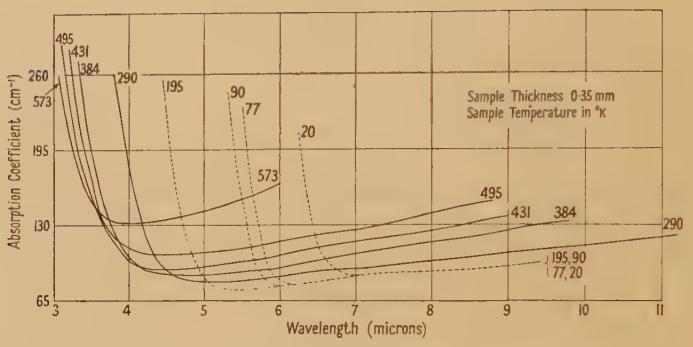
Absorption of p-type PbSe crystal (Gibson 1952).

For the laboratory-grown crystals no values of the minimum absorption coefficient as low as that obtained by Paul, Jones and Jones (1951) were found. The minimum value varied between about 30 and 300 cm^{-1}. Many samples of natural galena shew no observable transmission at all so that the sample found by Paul, Jones and Jones (1951), is clearly exceptional (Paul and Jones 1953).

Typical absorption spectra obtained by Gibson for a wide range of temperatures are shewn for PbS, PbSe, PbTe in figs. 8, 9, 10 respectively.

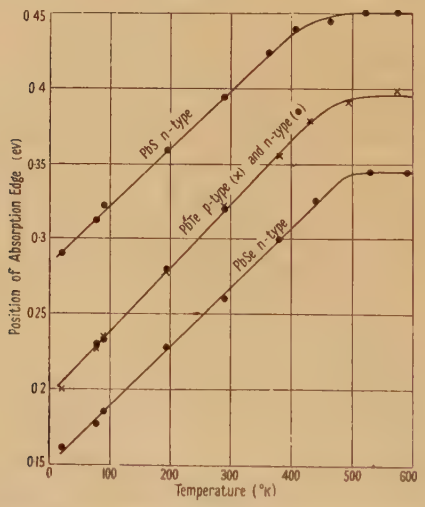
The common features of these are as follows. A very rapid fall in absorption occurs at a certain wavelength (absorption edge) which moves to longer wavelengths as the temperature is lowered except at temperatures above 500°K at which the variation is small. The rate of change in each case is almost the same as that found for the long-wave limit of photoconductivity and corresponds to an energy change of about 4×10^{-4} ev/°C. This variation is shewn in fig. 11.

Fig. 10



Absorption of n-type PbTe crystal (Gibson 1952).

Fig. 11



Position of absorption edge at various temperatures (Gibson 1952).

At a wavelength slightly longer than that corresponding to the edge the absorption passes through a minimum and then increases slowly but steadily with wavelength. In this region the absorption is approximately proportional to the square of the wavelength and increases as the temperature increases. As we have seen, the value of the minimum

absorption varies widely from sample to sample but Gibson finds no correlation between number of carriers and minimum value. There is a tendency for p-type samples to absorb more strongly than n-type samples but the difference is not very marked. In contrast, the position of the absorption edge appears to be quite characteristic of the material and is apparently the same for all samples n-type and p-type alike. This is a very important observation as it also appears to indicate that the photo-conductive limit is a characteristic of the material and not influenced by impurities such as oxygen as had been thought (see § 2.6). This follows from the observed correspondence of absorption edge and photo-conductive ' limit '.

The value of the absorption coefficient on the short-wave side of the absorption edge is difficult to measure but it rises rapidly to values of the order of 10^4 cm^{-1}, another strong indication that the absorption is characteristic of the basic substance and not an impurity absorption. For the latter, impurity concentrations greater than 0·1% would be required and these are most unlikely. So far, it has proved impossible to study the variation of the magnitude of the absorption in this region with carrier concentration. The absorption ' edges ' observed at considerably shorter wavelengths ($\sim 1\,\mu$ for PbS) with thin films cannot be observed with these samples as they cannot be made thinner than 0·05 mm. A few very thin small crystals of PbTe ($\sim 1\,\mu$) have been grown by Lawson by evaporation and Gibson (1952 c) has used these to measure the absorption down to about $1\,\mu$. The rapid rise in absorption between $1\,\mu$ and $2\,\mu$ found with evaporated layers and an absorption coefficient of about 10^5 cm^{-1} at $1\,\mu$ was again found. It would therefore appear that up to about $2{\cdot}5\,\mu$ the behaviour of layers and single crystals is similar. At longer wavelengths, however, the behaviour is quite different, the poly-crystalline layers having an additional high absorption of the order of 10^4 cm^{-1} extending out well beyond the photo-conductive limit at room temperature. Evaporated layers of thicknesses up to 0·05 mm have been prepared and behave quite differently from thin flakes of similar thickness obtained from a single crystal.

A different approach to the measurement of absorption in single crystals of these substances has been made by Avery (1951, 1952, 1953). By measuring the reflection coefficient at a plane crystalline surface at various angles of incidence for two planes of polarization he has been able to calculate the real and imaginary parts of the dielectric constant and hence to deduce the absorption. This method is not applicable for small values of the absorption such as occur for wavelengths greater than the photo-conductive absorption edge but is well adapted to give values of the absorption at shorter wavelengths for which the transmission method fails on account of the high absorption. These measurements also shew that for single crystals of PbS the absorption increases rapidly from a few times 10^4 cm^{-1} to over 4×10^5 cm^{-1} as the wavelength is reduced below $1\,\mu$. As the wavelength is increased it remains of the order of 10^4 cm^{-1} out to at

least $2{\cdot}5\,\mu$. Absorption coefficients actually somewhat higher than those given by Gibson (1952 c) are found. This is not surprising in view of the difficulty in obtaining uniform layers. Similar results are also obtained for PbTe.

From Avery's measurements values for the refractive index may be found. Only a small variation is found between $1{\cdot}5\,\mu$ and longer wavelengths up to $3{\cdot}5\,\mu$. For shorter wavelengths the value begins to decrease. At a wavelength of $3\,\mu$ the average values for the refractive index n for a number of specimens are given as follows.

	n at $3\,\mu$
PbS	$4{\cdot}01 \pm 0{\cdot}06$
PbSe	$4{\cdot}59 \pm 0{\cdot}06$
PbTe	$5{\cdot}35 \pm 0{\cdot}10$

It will be noted that the refractive index follows the natural order and not that of the photo-conductive and absorption edges.

The bearing of these results on the determination of the band structure of these crystals will be discussed in § 6.

3.4. *Photo-voltaic Effects and Photo-conductivity in Single Crystals of PbS, etc.*

It has long been known that a photo-voltage may be produced at a point contact between a metal and some PbS crystals and that the spectral characteristic is similar to that for photoconductivity in layers (see Elliott 1947, Sutherland and Lee 1948). It is also well known that such contacts frequently shew marked rectifying action (see § 3.5).

Photo-conductivity is also shewn by such contacts under certain conditions. This generally manifests itself through a reduction of the resistance in the 'back' direction of a rectifying contact on illumination. It is not easy to separate the photo-voltaic and photo-conductive effects but an attempt to do this has been made by Gibson (1952 a, b) in a detailed study of the photo-conductive properties of such contacts with single crystals of PbTe and PbS. For PbTe marked rectifying properties are only found with p-type crystals and then only when the temperature is reduced to about 190°K. For such contacts at 90°K marked photo-conductive effects are found and also a photo-voltaic effect. The photo-conductive effect is only appreciable in the 'back' direction. For n-type crystals only a small effect is found and only after treatment which may well produce a p-type layer at the surface. Similar results are found both with crystals of natural galena and laboratory grown single crystals of PbS. Again only p-type crystals shew marked photo-conductive effects. For PbS, however, the effect occurs quite readily at room temperature. This is in line with experience with photo-conductive layers. The spectral response in both cases is almost identical with that for a well-sensitized photo-conductive cell.

Gibson has interpreted his observations in terms of a Schottky-type barrier which is created at the surface of the semi-conductor in contact with

the metal point. The resistance changes observed are assumed to be due to a change in the effective height of the barrier by the incident radiation. This leads to a reasonable interpretation of the variation of photo-current with applied voltage. For PbTe an effective barrier height of the order of 0·1 ev is found at 90°K decreasing to a small value at room temperature. For PbS the value at 90° is about 0·2 ev and 0·1 ev at room temperature.

By illuminating the contact with a small spot of light the effective range of the carriers which produce the photo-effect was measured. This varies from sample to sample but distances of the order of 50 μ for the signal to fall to half its peak value were found.

Similar behaviour was later found for PbSe and, as we have seen (§ 2.6.3) it was a measurement of the photo-conductivity at a point contact with a PbSe crystal that first established the correspondence for this substance between the photo-conductive limit and absorption edge and shewed that the long wave limit for PbSe lies well beyond that for PbTe (Gibson, Lawson, and Moss 1951).*

3.5. *Rectification and Transistor Action in Single Crystals of PbS, etc.*

It has long been known that some galena crystals shew strong rectifying action and such crystals were widely used as detectors in the early days of wireless. More recent studies have shewn that only p-type samples of PbS shew marked rectification (Legrand 1948, Arianova, and Sokolskaya 1951) and this has also been verified using synthetic crystals of PbS, PbSe, and PbTe (Gibson 1952 a, b). Current-voltage characteristics for PbS have been studied by Henisch and Granville (1951). The variation of rectifying properties by various surface treatments has been discussed by Hogarth and Granville (1951) and the effect of introducing various constituents into the bulk crystal, including excess sulphur by Riemann and Sullivan (1952).

In view of the similarity of this group of substances with germanium, as far as their properties as semi-conductors are concerned, it would seem reasonable to expect them to shew transitor action, i.e., the influence of the current at a contact biassed in a ' forward ' direction on the current at a nearby contact biassed in the ' back ' direction (see e.g. Shockley 1951). This was first observed for p-type samples of PbS by Gebbie, Banbury and Hogarth (1950) (Banbury, Gebbie and Hogarth 1952). Voltage gains of the order of 25 are common and even higher values are sometimes found but, so far, current gains greater than unity have not been reported. The frequency characteristics of such transistors have been studied by Banbury and Henisch (1950) who have shewn that a rapid fall off in response occurs for frequencies above about 200 kc/s. That the transistor effect is not wholly a surface phenomenon has been shewn by Banbury (1952) by placing the ' emitter ' and ' collector ' on

* *Note added in proof*:—Since this was written bulk photo-conductivity has been observed in single crystals of PbS by T. S. Moss. The author is grateful to Dr. Moss for this information prior to publication.

opposite sides of a thin crystal. By this means he has also been able to measure the mean range of the 'injected' carriers. For the materials used the value is of the order of 50μ which is also the value found by Gibson (1952 b) for optically injected carriers from his study of the photo-conductivity of point contacts with PbS crystals.

Transistor action has also been found by Hogarth (1953) for laboratory-grown (T.R.E.) crystals of PbSe. The effects are not so marked as for the best PbS specimens but this is almost certainly due to the fact that the synthetic crystals are not so pure as the best samples of natural galena. Transistor effects have also been found by Hogarth (1953) for PbTe but only at reduced temperatures (90°K). This is in accord with Gibson's (1952 a) observations on the photo-conductivity of point contacts with PbTe. Again the effects are only marked for p-type samples of all three substances.

3.6. *Other Properties of the PbS Group of Semi-conductors*

We shall discuss in this section various other properties of this group of semi-conductors. They all have a sodium chloride crystal structure, the edge of the unit cubic cell being as follows:

PbS	5·97 A.U.
PbSe	6·14 A.U.
PbTe	6.45 A.U.

It will be seen that these follow the natural order.

Accurate measurements have recently been made of the molecular heats of all three substances of the group for temperatures between 14°K and 300°K by D. H. Parkinson* at T.R.E. It is found that, over this temperature range there are no specific heat anomalies and that the values of the molecular heat increase smoothly with temperature, that of the selenide lying between those of the sulphide and telluride over the whole range. This is most interesting in view of the behaviour of the electronic properties previously discussed. We may also note that recent measurements of the heats of formation also indicate that these follow the natural order (National Bureau of Standards, 1948).

Lead sulphide has been thought for some time to become superconducting at low temperatures and this seemed to be confirmed for all three substances of this group by more recent measurements by Darby, Hatton and Rollin (1950). This is somewhat surprising in view of their other properties as semi-conductors. Using moderately pure samples of p-type natural galena (7×10^{17} and 2×10^{16} carriers per c.c. at room temperature) in later measurements Hatton, Rollin, and Seymour (1951) have found no evidence of super-conductivity down to a temperature of 1°K. They conclude that previous observations of superconductivity may have been due to formation of lead filaments at grain boundaries in the

* The author is indebted to Dr. Parkinson for use of this information prior to publication.

samples. This conclusion is also reached by Hudson (1951) who found no evidence of super-conductivity except for specimens for which x-ray examination shewed the presence of lead. In particular no evidence of super-conductivity was found for a synthetic crystal of PbTe (T.R.E.) down to a temperature of 1·3°K. It may therefore be concluded that these substances in pure crystalline form are not superconducting, at least in the ' helium ' range of temperatures.

Thermo-electric power measurements have been made for single crystals and an attempt to correlate the direction of rectification, the sign of the photo-voltaic effect and thermo-electric power has been made by Granville and Hogarth (1951). It is found that a correlation only exists for freshly cleaved surfaces or for surfaces which have been carefully etched after polishing. It is clear that great care must be taken in the interpretation of thermo-electric power measurements if used to give an indication as to whether a bulk sample is n-type or p-type.

§ 4. Other Infra-red Photo-conductors

A large number of compounds and one or two elements are photo-conductive in the near infra-red between 1 μ and 2 μ. There are very few however, which have their $\lambda_{1/2}$ value (see § 2.6.1) greater than 2 μ.

4.1. *Compounds Involving S, Se, or Te*

A number of such compounds has been discussed by Moss (1950 a, 1952 b). None of these have $\lambda_{1/2}$ greater than 2 μ, although some, and in particular MoS_2 for which $\lambda_{1/2}=2\,\mu$, may exhibit photo-conductivity at wavelengths greater than 2 μ. The photo-conductivity of a number of such compounds has been investigated by Braithwaite (1951) using evaporated layers. His values for $\lambda_{1/2}$ and λ_m, the maximum wavelength at which photo-conductivity has been observed are given below, when the latter exceeds 2 μ.

Material	$\lambda_{1/2}$ (μ)	λ_m (μ)
Cu_2 Te	1·3	2·2
Ag_2 Te	1·3	3·0
Zn Te	1·4	3·5
Hg Te	3·1	3·9
Tl_2 Te	1·6	2·6
Sb_2 Te_3	1·6	2·6
Mo Te_2	1·6	2·6
U Te_2	1·35	2·5

It should be noted that the above values of $\lambda_{1/2}$ and λ_m were measured at 90°K. Of these substances it will be seen that only Hg Te compares at all with the PbS group.

The photo-conductivity of the series of bismuth compounds which corresponds to the PbS group has been studied by Gibson and Moss (1950) using evaporated layers. They behave similarly in many respects, the

$\lambda_{1/2}$ value shifting to longer wavelengths as the layer is cooled. Their sensitivity is greatly increased by oxygen treatment but does not reach such high values as found for the PbS series. The sulphide shews very little sensitivity at wavelengths greater than 2μ but the values for $\lambda_{1/2}$ and λ_m for the telluride at 90° K are about $2{\cdot}8\mu$ and 4μ respectively. As for the lead series the telluride shews no sensitivity at room temperature whereas the sulphide does.

A large number of compounds have been studied by Schwarz (1948) as possible materials for use in photo-cells. Of these only InTe appears to be sensitive beyond 2μ and this only as far as $2{\cdot}2\mu$. We may note, however that CdSe shews very high sensitivity (Schwarz 1950), comparable with that for the PbS group, but is limited to wavelengths less than 2μ.

4.2. *Intermetallic Compounds*

The intermetallic compound Mg_3Sb_2 has been shewn to be photo-conductive in the infra-red by Zhuse, Mochan, and Ruirkins (1948) and by Moss (1950 b). Layers evaporated in the pure state were found by Moss to shew no photo-conductivity at wavelengths longer than $2{\cdot}8\mu$ but when evaporated in a low pressure of air the value of $\lambda_{1/2}$ moved out to $3{\cdot}5\mu$.

None of these substances have been studied in anything like the detail with which the PbS group has been. In particular, work with single crystals is lacking. It would clearly be of great interest to make a comparable investigation with some of these, particularly HgTe, Bi_2Te_3 and intermetallic compounds such as Mg_3Sb_2 which shew photo-conductivity well beyond 2μ.

4.3. *Photo-conductivity in Elements*

A number of elements which behave as semi-conductors or insulators shew photo-conductivity and several of these have $\lambda_{1/2}$ values in the near infra-red. These have been fully discussed by Moss (1951, 1952 b). Of the semi-conductors those which have been most extensively studied are silicon and germanium with $\lambda_{1/2}$ values of $1{\cdot}1\mu$ and $1{\cdot}7\mu$. Germanium is available as single crystals of exceptionally high purity and it is interesting to compare the properties of this substance with those of the PbS group. This we defer till the next section.

The only element which exhibits photo-conductivity at ordinary temperatures at wavelengths appreciably greater than 2μ is tellurium. The bulk properties of this substance, which behaves as a semi-conductor when pure, have been fairly extensively studied and its photo-conductivity has been observed by the use of evaporated layers by Moss (1949 a, 1952 b). A value of $\lambda_{1/2}$ of about $3{\cdot}7\mu$ is obtained for layers cooled with liquid air and for λ_m about $4{\cdot}3\mu$. Very little photo-conductivity is observed at room temperature. The value of $\lambda_{1/2}$ moves to longer wavelengths as the temperature is reduced (Moss 1950 c) the shift corresponding to about 2×10^{-4} ev/°c.

Tellurium has been studied as a semi-conductor by a large number of workers but, as has usually happened, until fairly recently samples of sufficient purity were not available. Bottom (1948, 1949) established that, in its pure state, tellurium is a semi-conductor shewing intrinsic conductivity at room temperature. His value for the forbidden energy gap ΔE is 0·38 ev. Johnson (1948) found a similar value for ΔE, and room temperature mobilities, derived from Hall constant measurements, of about 550 cm/sec per volt/cm for both holes and electrons. Extensive measurements using highly purified material in the form of single crystals by Fukuroi, Tanuma, and Tobisawa (1949 a, b; 1950 a, b) have given a value 0·34 ev for ΔE and room temperature mobilities $\mu_e = 1600$ cm/sec per volt/cm $\mu_h = 1100$ cm/sec per volt/cm. The variation of mobility with temperature is somewhat complex owing to the anisotropic nature of tellurium and for certain ranges $\mu_h > \mu_e$. This leads to double reversals of the sign of the Hall constant.

Measurements by Moss (1952 b) have shewn that an increase in absorption occurs at wavelengths less than $3{\cdot}5\mu$. For longer wavelengths the absorption is constant and of the order of 130 cm^{-1}. Loferski and Miller (1951) have given a value $4{\cdot}2\mu$ for the position of the absorption edge for their samples, which were of high purity. Absorption coefficients about ten times smaller than that given by Moss were found on the long wavelength side of the edge. The absorption edge therefore coincides well with the long-wave limit of photo-conductivity.

It is of interest to note that a high value of refractive index was found by Moss (1952 b) who gives a value of approximately 5 for wavelengths greater than 4μ. It will therefore be seen that in many respects tellurium behaves like the PbS group of substances.

Measurements on grey tin by Kendall (1950) and by Busch, Wieland, and Zoller (1951) have established that it is a semi-conductor in its pure state with energy gap ΔE of the order of 0·1 ev. Photo-conductivity has not yet been observed, but should it be it would be expected to extend to a wavelength beyond 10μ. A form of antimony has been prepared by Moss (1952 a, b) as evaporated layers which behaves as a semi-conductor with ΔE in the range 0·05–0·2 ev. Variation in resistance on illumination with infra-red radiation at wavelengths up to 16μ has been observed but it has not been established whether this is a bolometric effect or photo-conductivity.

4.4. *Photo-conductivity at Very Low Temperatures*

At very low temperatures an effect of great interest has been observed. Rollin and Simmons (1952, 1953) have shewn that bulk crystalline samples of silicon become photo-conductive at wavelengths in the range $2–14\mu$, at liquid hydrogen temperatures. Measurements were not made at longer wavelengths but there is evidence that the photo-conductive effect extends well beyond 14μ. This effect was interpreted as due to excitation of electrons from impurity levels. The samples used had

10^{16}–10^{17} impurity centres with energy levels lying about 0·05 ev above the full band. If the above interpretation is correct one would expect the long-wave limit of this photo-conductivity to be at about 25μ. It seems likely that similar effects will be observed in other semi-conductors. They are likely to occur only at low temperatures as otherwise such acceptor centres will all be filled.

§ 5. Comparison of PbS Group with Ge and Si

A great deal of work has been done on the electrical and optical properties of Ge and Si—particularly the former. Ge may now be obtained so pure as to shew intrinsic conductivity at room temperature, and the amounts and kind of impurity may be controlled. It is therefore an ideal substance to act as a standard of comparison for other semi-conductors. Also, being an element, its behaviour would be expected to be simpler than for compound semi-conductors such as PbS. The electrical and optical properties of Ge and Si have been described by Shockley (1951) and work on their photo-conductivity has been summarized by Moss (1952 b).

The electrical properties of Ge and Si and the PbS group are very similar in many respects. The value of ΔE for Ge is 0·74 ev and for Si is 1·1 ev. The properties of Ge and Si should therefore be expected to resemble those of PbSe, PbTe and PbS (see § 3.2). This is in fact so but there are important and interesting differences. PbTe and PbSe should be intrinsic conductors at room temperature like Ge but the laboratory grown single crystals available are not yet of sufficient purity to shew this. All have high electron and hole mobilities, but for Ge life-times of minority carriers are very much greater than for the PbS group for which, so far, values have not been found to exceed a few microseconds. This again may be largely a question of purity. Ge shews marked rectifying and transistor action when n-type whereas Si and all three of the PbS group do this when p-type. The most important difference between the two groups of substances is, however, that whereas for Ge and Si the long-wave limit of strong absorption and photo-conductivity correspond to energies which are very close to that of the forbidden energy gap as determined by conductivity and Hall constant measurements in the intrinsic region this is not so for the PbS group. Thus in Ge and Si and also, for example, in Te there is no doubt that the photo-conductive limit corresponds to the transition of an electron from the top of the filled band to the bottom of the conduction band. This cannot be so for the PbS group unless the forbidden gap ΔE is greatly reduced at ordinary and low temperatures. For example in PbS the value of ΔE for the intrinsic range is 1·17 ev whereas the quantum energy corresponding to $\lambda_{1/2}$ is about 0·4 ev.* The latter also corresponds to the energy of a quantum absorbed at the long-wave absorption edge. For Ge and Si, like most photo-conductors, the long-wave limit of photo-conductivity moves to shorter wavelengths as the temperature is reduced,

* See footnote p. 341.

whereas for the PbS group the opposite behaviour is observed. Te however behaves like the PbS group in this respect. These differences will be discussed further in the next section.

§ 6. Theory of Photo-Effects in the PbS Group of Substances

In developing any theory of the marked photo-effects in the infra-red which characterize the PbS group of substances some model must be assumed to represent the basic semi-conductor. We may assume that the forbidden energy gap is given by Putley's Hall constant measurements with single crystals (see § 3.2) and there is very little doubt that this is so in the range of temperatures (above 500°K) at which intrinsic conductivity takes place.* This assumption has, however, not been made by many of those who have discussed the theory of photo-conductive layers. They have supposed that the energy corresponding to the long wave limit of photo-conductivity is equal to the energy required to bridge the forbidden gap. Unless some change in the electronic structure of PbS, etc., occurs at temperatures below the intrinsic range it is most unlikely that the energy gap would decrease to this extent, for example in PbS, from 1·17 ev at 500°K to about 0·4 ev at 300°K. (It is known that no marked change in crystal structure takes place.) The rate of decrease at lower temperatures corresponding to the shift of the photo-conductive limit would be about 4×10^{-4} ev/°C (see § 2.6). Also Gibson's (1952 c) measurements on the absorption of single crystals have shewn that the rate of shift of the absorption edge with temperature decreases rather than increases above room temperature (fig. 11). It is important to appreciate this difficulty in discussing some of the theoretical work on the photo-effects.

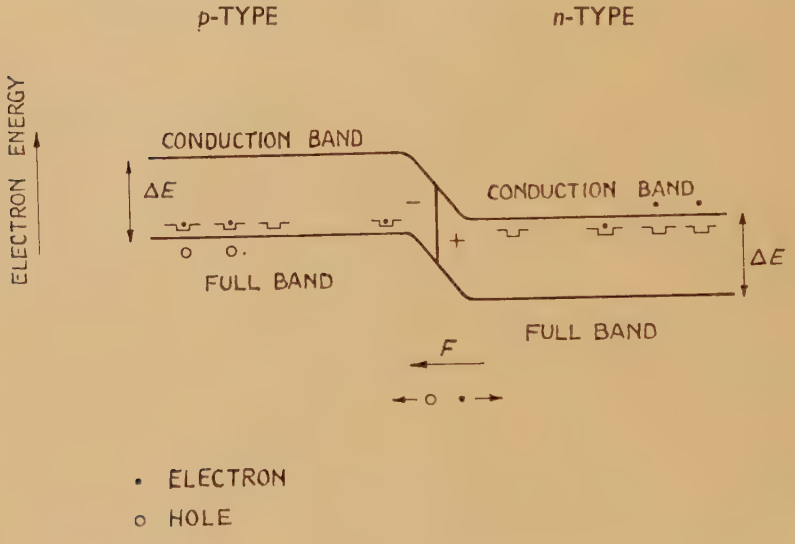
6.1. *Theory of Photo-effects in Layers*

The first attempt to give a theory of the extremely high photo-sensitivity of evaporated layers of PbS was that of Sosnowski, Starkiewicz and Simpson (1947). From their experimental work, as discussed in § § 2.4, 2.6, they concluded that the layers were polycrystalline and that the main part of the photo-effect was due to the action of the radiation on intercrystalline barriers. The evidence for this is very strong and has already been reviewed in § 2.4. They supposed that the intercrystalline barriers were due to p-n junctions between the micro-crystals making up the layer. This was deduced from the observation that the resistance of the layer reaches a maximum when an n-type layer is just turned to p-type by baking in oxygen. They concluded that in this condition some micro-crystals would be p-type and some n type. The donor centres in the n-type crystals were assumed to be due to excess lead and the acceptor centres in the p-type to excess oxygen. The theory of a contact between an n-type and p-type crystal of PbS has been discussed by Sosnowski (1947) and is now well known for germanium (see Shockley 1951). The photo-effects obtained on illuminating such a junction have

* See footnote p. 341.

been fully discussed, particularly for germanium, by Becker and Fan (1950). The elementary principles are illustrated in fig. 12. Electrons from the donor centres near the surface of the n-type material flow into acceptor centres near the surface of the p-type material. Thus a positive space charge is created in the n-type material and a negative space-change in the p-type. This proceeds till the electric potential in the n-type material is raised sufficiently to bring the level of the top of the full band in the p-type material approximately to the level of the bottom of the conduction band in the n-type material. Thus a potential barrier is set up between the two of height approximately equal to the width ΔE of the forbidden energy gap and a strong electric field is produced at the junction. Hole and electron pairs created by radiation near the

Fig. 12

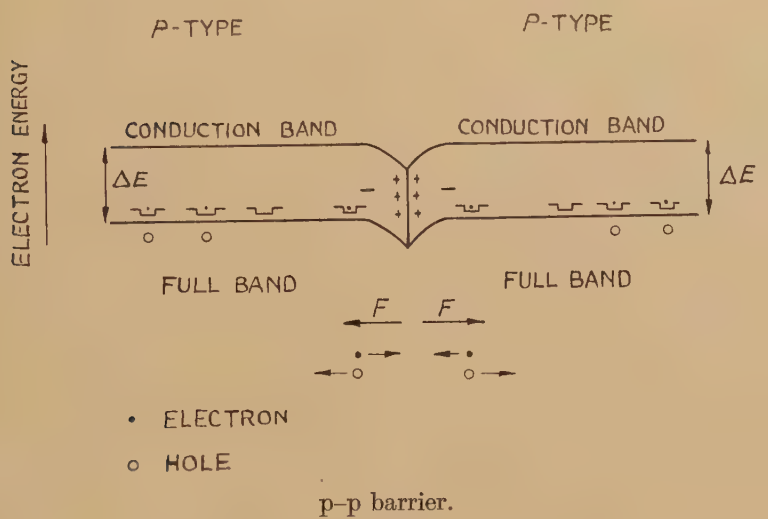


p–n barrier.

junction will thus be separated and will change the equilibrium charge distribution, thus producing a photo-voltage. Photo-conductivity is supposed to arise through modification of the barriers by the free charges created. The effect of space charge barriers has been further discussed by Schwarz (1949). The above model has also been extended by James (1949) and Rittner (1950) to include variations in impurity concentration causing changes from p-type to n-type within the bulk material itself. These authors suppose that the primary photo-effect is the excitation of an electron from the full to the conduction band, an energy of about 0·4 ev being required for this in the case of PbS. If indeed the energy required is 1·17 ev as might be supposed from Putley's measurements, the potential barriers which would be created would be very high and some other source must be sought for the primary photo-electrons.

In order to avoid this difficulty it has been suggested by Smith (1951) that a model previously proposed by Mitchell and Sillars (1949) to explain certain effects in silicon carbide might be applicable here. This, he discussed in terms of n-type material but it is now known that photo-effects mainly occur in the PbS group with p-type material. The essentials of the model still apply, however. Surface states which trap holes can create a space charge barrier in p-type material whose height can vary over a wide range of values according to the density of surface states and the properties of the material. This is illustrated in fig. 13. Any photo-electrons created near such a barrier would be sucked into it and would release some of the trapped holes. This in turn would reduce the height of the barrier. Any asymmetry between two crystals in contact would create a photo-voltage as before. Photo-conductivity is thus largely

Fig. 13



p–p barrier.

supposed to be a 'tap' effect, the primary photo-electrons merely reducing the height of the barriers and letting more current flow. This model has been discussed in detail by Gibson (1951) who has given the elements of its theory and has described a number of experiments which may be interpreted as shewing its validity. In particular, effective heights of the barriers may be deduced. These come out on the average to be not more than about 0·2 ev which is much lower than one would expect from a p-n junction. Barriers of similar height have been found at the contacts between a p-type crystal and a metal contact (see § 3.4). It therefore appears that although p-n junctions may play some part in these photo-effects, p-p junctions are likely to be more important.

Gibson (1951) has made a comparison of this theory involving barriers with that assuming the incident radiation simply to increase the number

of electrons in the conduction band or the number of holes in the full band. He concludes, particularly from the variation of time-constant of the photo-conductive effect with temperature and with intensity of illumination, that the barrier theory only can explain the effects.

Simpson and Sutherland (1951) however, have interpreted their results with PbTe in terms of a simple bimolecular recombination process and have concluded that, provided it can be assumed that there is a 'tail' in the distribution of energy levels in the full or conduction band, no barriers are required. They suggest that the barrier picture only applies to layers treated with oxygen. They tried to exclude oxygen from their layers as far as possible. It may well be that both bulk changes in conductivity and barrier effects are present as has recently been suggested by Ewald (1951) in order to explain the photo-conductivity of Tl_2S. It seems fairly clear, however, that the barriers play a predominant part in layers sensitized with oxygen.

6.2. *Source of Primary Photo-electrons*

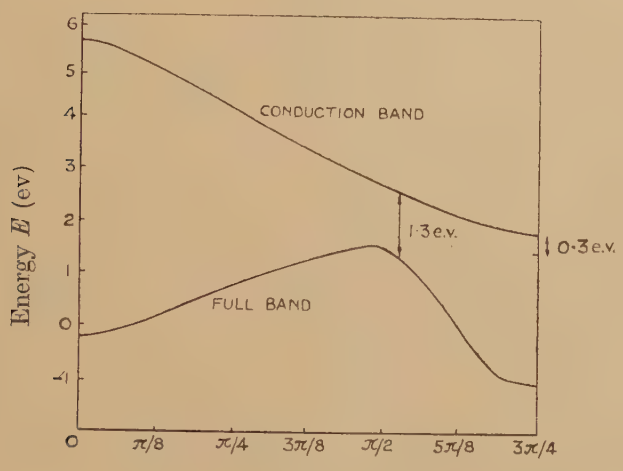
The various theoretical treatments discussed above really throw no light on the source of the primary photo-electrons. All the theories simply depend on the creation of hole-electron pairs and work equally well however they are created. Most of the authors, apart from Gibson (1951), assume that the primary photo-electrons come from a transition from the full to the conduction band. We have already pointed out the difficulties of this assumption. Various impurity centres have also been considered as possible sources of the primary photo-electrons. For example Pincherle (1951) has considered F-centres, F′-centres and F″-centres, i.e., sites from which a negative ion is missing and at which are trapped one, two or three electrons. He concludes that the F and F′ centres are too deep and that F″ centres are unstable in PbS. He has also discussed D-centres, i.e., pairs of missing neighbouring positive and negative ions. Such centres give levels (about 0·25 ev) below the conduction band and may play an important part as electron traps, but are unlikely to be the source of the primary photo-electrons. Indeed it seems clear from the high value of the absorption coefficient ($\sim 10^4$ cm $^{-1}$), in single crystals, to the short wave side of the photo-conductive edge, that impurity centres cannot be the source of the primary photo-electrons. We are then left with two possibilities: (1) that the forbidden gap is indeed of the order of 0·4 ev for PbS, and less for PbTe and PbSe, at room temperature, or (2) that electrons are first excited to states of lower energy, characteristic of the basic lattice, but insufficient for the creation of free holes, i.e., that excitons are formed (see Mott and Gurney 1940, p. 84). Pincherle (see § 6.3) has recently calculated the energy required to create such an exciton in PbS as about 0·3 ev, the right order of magnitude. It must still be explained how the exciton is broken up to form a hole-electron pair, a process requiring a further 0·9 ev of energy, if we assume Putley's value for the forbidden energy gap ΔE. It must

therefore be admitted that no satisfactory answer is at present available as to the source of the primary photo-electrons, and this is the outstanding problem of the subject.

6.3. *Theory of Electronic Band Structure of PbS*

An attempt to solve the difficult problem of calculating by means of quantum mechanics the electronic energy levels for the PbS group of semi-conductors has been made by L. Pincherle and his colleagues at T.R.E. (Bell *et al.* 1953). The so-called ' cellular ' method is used, self-consistent ' radial ' wave-functions being calculated for each elementary cell into which the lattice can be divided, and fitted across the boundaries. The self-consistent calculation is only carried out for zero value of $\boldsymbol{k}$, the momentum vector, and this turns out to be the chief weakness of the

Fig. 14



E–k curves for (1, 1, 0) direction of PbS crystal.

calculation. Preliminary results shew that the full and conduction bands are separated by a small gap, which does not occur at the edge of a Brillouin zone as expected. The value of the gap comes out as about 0·3 ev, but the calculations in their present form can only give the order of magnitude of ΔE. This value cannot therefore be taken as strong evidence against Putley's higher experimental value 1·17 ev (see § 3.2). The form of the energy levels as a function of the momentum vector $\boldsymbol{k}$ for one direction in the crystal is shewn in fig. 14 for PbS. The two curves shewn are for the conduction band and the highest level in the full band, which has two lower overlapping levels.

It will be seen that the full band has a maximum for a certain value of $\boldsymbol{k}$. If this maximum were a sharp ' peak ' it might explain the low value of the energy associated with the long-wave optical absorption

edge as compared with the value of ΔE found by Hall constant measurements. The shape of this maximum is sensitive to small changes in the electronic wave-functions and it will be most interesting to have the results of more accurate calculations.

Using the self-consistent field obtained from the above calculation, the energy required to form an exciton has been calculated.* This turns out to be about 0·3 ev. This value is much less sensitive to the form of the wave-functions than the energy gap ΔE and is unlikely to be far wrong. It will be seen that it is much greater than the value which would be obtained by calculating the energy levels in a Coulomb field with modified dielectric constant—a method sometimes used in the literature to estimate the energy of formation of an exciton. This energy of the first excited state above the ground level is given by $10{\cdot}1/n^4$ ev where n is the appropriate refractive index. With $n=4$ this gives 0·04 ev.

6.4. *Relationship between Long-wave Limit and Dielectric Constant*

It has been shewn by Moss (1950 a, 1952 b) that there exists an approximate relationship between the dielectric constant n and the $\lambda_{1/2}$ value for the long-wave 'limit' of photo-conductivity for many photo conductors in the form of compounds and including the PbS group. This is

$$n^4/\lambda_{1/2}=77, \quad \ldots\ldots\ldots \quad (4)$$

where $\lambda_{1/2}$ is expressed in microns. Moss has also discussed possible explanations of this form of relationship. If one assumes that the long-wave limit of photo-conductivity is set by impurity centres with a Coulomb type of field, a relationship like (4) would hold provided that the same kind of centre were active in all photo-conductors for which the relationship holds. Pincherle (1951) has shown that for the F-type of centre the Coulomb field is a poor approximation and in any case it is unlikely that the same kind of centre will be responsible for the photo-conductivity for the wide range of substances for which the relationship holds approximately. Some such relationship must hold in view of the connection between refractive index and absorption but it is unlikely that so simple a relationship could be deduced from the complex form of absorption observed. At present it seems best to regard (4) as an empirical relationship, useful in deciding approximately the position of the long-wave limit of photo-conductivity of a substance whose refractive index is known. It should be stressed that the relationship is only approximate and indeed gives the wrong order for the limits of PbTe and PbSe.

§ 7. Photo-conductive Cells as Infra-red Detectors

The properties and use of PbS photo-conductive cells as infra-red detectors have been discussed in their reviews by Elliot (1947) and by Sutherland and Lee (1948). The former deals largely with German work during the war and the latter also covers this work but includes British

* I am indebted to Dr. L. Pincherle for this information prior to publication.

and American work up to 1947. The detectors described in these reviews have been largely superseded by modern practice and in particular much higher sensitivities without cooling have been obtained both with evaporated-layer PbS cells and with cells using chemically deposited layers. In the former, the sensitive layer is normally enclosed in an evacuated glass envelope whereas, in the latter, it is generally deposited on a sheet of glass or mica and simply protected from the atmosphere by a thin film of transparent material. Vacuum mounting is, however, also used for chemical cells.

Surprisingly little has been published on the actual performance of photo-conductive cells as infra-red detectors. This also applies to details of manufacture which still appear to involve a certain amount of laboratory ' skill '. Various methods of preparation of photo-conductive layers have already been discussed in § 2.6 and references to the relevant papers given. In a recent review Simpson and Sutherland (1952) have given a certain amount of data on recent American practice and Milner and Watts (1952) and Moss (1952) have reviewed some recent British work. We shall discuss in turn such information as is available on PbS, PbTe, and PbSe cells and shall follow this with a brief discussion on the fundamental limits to the sensitivity of such cells. We shall deal only with photo-conductive cells since, so far, no other type comes near to them in ultimate sensitivity.

Before proceeding to this discussion of sensitivity we must make a distinction between two quantities which are relevant. These are the sensitivity of response of a cell and the minimum energy which it can detect. The former we shall call the ' responsivity ' and the latter the ' ultimate sensitivity '. It does not follow that a cell with the highest responsivity will be able to detect the smallest radiant energy as the ultimate sensitivity is conditioned by the noise level of the cell. This is usually in the form of current noise due to the exciting current through the cell.

Suppose we have a cell of resistance R connected to an amplifier through an equal load and excited by a voltage V as shewn in fig. 15. Then the signal voltage V_s produced by a change of resistance ΔR will be given by

$$V_s = V\Delta R/4R \qquad (5)$$

provided $\Delta R \ll R$. Now if an amount of incident energy W causes a change in resistance ΔR we may write, for small values of W, $\Delta R = \alpha W$. We then have

$$V_s = V\alpha W/4R = rW. \qquad (6)$$

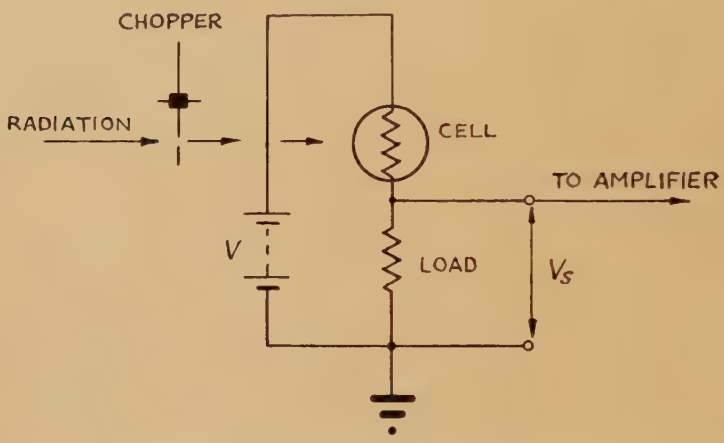
The quantity r is generally called the ' responsivity ' and is expressed in volts per watt. It depends on the exciting voltage V.

If V_n is the r.m.s. noise voltage appearing across the load the voltage signal-to-noise ratio will be given by V_s/V_n. In particular when this is unity the value W_m of W will be given by

$$W_m = V_n/r. \qquad (7)$$

This value is conventionally taken as the minimum detectable energy. Normally for current noise V_n is approximately proportional to V so that to a first approximation W_m is independent of V. This is usually found to be so over a fairly large range of V, which is made sufficiently large to make current noise predominant but small enough not to heat the layer appreciably. For current noise in a given type of layer of area A, V_n is proportional to $A^{1/2}$ for a given current so that a small area cell may be expected to shew a smaller value of W_m as compared with a similar cell of larger area. This makes comparison of actual cells somewhat difficult, and it is desirable, for this purpose, to reduce the value of W_m to that for a standard area. We shall take 0·1 cm² as this standard since this value is frequently used in practice (1 cm × 1 mm).

Fig. 15



Arrangement for use of photo-conductive cell.

The value of V_n also depends on the bandwidth Δf of the amplifier and display equipment. For values of Δf so small that the noise spectrum is effectively uniform over Δf, V_n is proportional to $\Delta f^{1/2}$. For current noise, however, over a wide range of frequencies from a few c/s to several kc/s the noise power per unit bandwidth is proportional to $1/f$ and the proportionality of V_n to $\Delta f^{1/2}$ may no longer hold for practical values of Δf. It is convenient to reduce the value of W_m to refer to a standard bandwidth, say 1 c/s, but care must be taken in interpretation of the value so obtained.

7.1. *PbS Infra-red Photo-conductive Cells*

Several authors (e.g. Strong 1951, Simpson and Sutherland 1952) have quoted PbS cells as being roughly 100 times better as regards ultimate sensitivity than a good thermocouple. This would lead us to expect to have a minimum detectable energy of the order of 10^{-12} watt for monochromatic

radiation for a 1 c/s bandwidth within the region of high sensitivity. This gives only the order of magnitude as no area is quoted. According to Moss (1953), for a PbS cell at room temperature 0·7% of the energy of a 200°C source is effective, so that the value of W_m for 200°C black-body radiation would be about 10^{-10} watt for a 1 c/s bandwidth, i.e., comparable with a thermocouple. Milner and Watts (1952) quote values for W_m of only $2{\cdot}5\times10^{-8}$ watt for 200°C radiation and 2×10^{-11} watt for 2·2 μ radiation for certain commercial cells. These therefore, appear to be somewhat lower in ultimate sensitivity than would be expected from the above. These values are reduced to 5×10^{-9} watt and 4×10^{-12} watt on cooling with solid CO_2. This brings them into the expected range. These values refer to a frequency of 800 c/s. On cooling, the time constant of the cells is increased from 40 to 200 μs.

Data on commercial chemically deposited PbS cells published by the Eastman Kodak Co. (1952) give values of W_m for 500°K radiation of the order 3×10^{-8} watt at room temperature when reduced to an area of 0·1 cm^2. The bandwidth, however, is not stated. The time-constant of these cells is rather long—of the order of 200–300 μs and the values of W_m refer to a frequency of 90 c/s.

It is clear from the experience of spectroscopists who have used PbS cells that values of W_m smaller than the above have been obtained with cells made under laboratory conditions. For such a PbS cell, for example, Fellgett (1949) quotes a value of W_m at maximum sensitivity of 8×10^{-13} watt/cycle (reduced to an area of 0·1 cm^2). This value was obtained by cooling with solid CO_2. This was an evaporated-layer cell and it appears that somewhat lower values of W_m, particularly at room temperature, may be obtained with this type of cell than with chemical cells. Experience of such laboratory-made cells would appear to indicate that with the best cells very little gain is obtained on cooling, apart from an extension of the long-wave limit. These cells have generally time-constants of the order of 20–100 μs at room temperature. The spectral response of a modern PbS cell is shewn in fig. 16. The variation with temperature is similar to that given by Moss (1947, 1949 b) (fig. 5). Such cells shew a remarkably constant response per quantum over a range of wavelengths from about 1 μ to 2·5 μ at room temperature before falling off with a $\lambda_{1/2}$ value of about 2·9 μ and reaching 1% of their maximum sensitivity at about 3·35 μ.

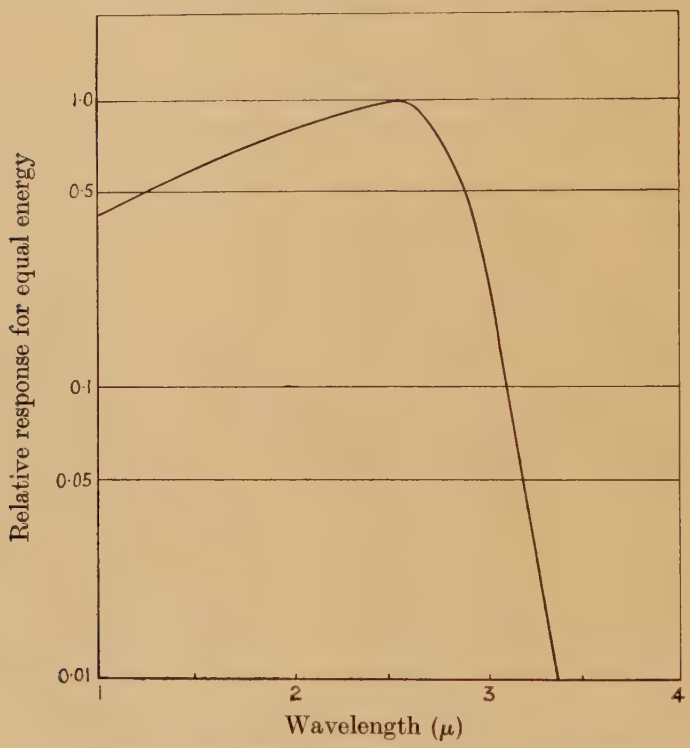
7.2. *PbTe Infra-red Photo-conductive Cells*

The main advantage of PbTe cells over those made with PbS is that considerably longer wavelengths may be reached by their use. There appears also to be some indication that slightly higher maximum sensitivities are also obtained in practice. Ultimate sensitivities of 100–1000 times that of a good thermo-couple have been quoted by various authors (e.g. Smith 1950, Simpson and Sutherland 1952). We should thus expect minimum detectable energies for monochromatic radiation of the order of 10^{-12} to 10^{-13} watt for a 1 c/s bandwidth. A figure 2×10^{-14}

watt has been quoted by Simpson, Sutherland and Blackwell (1948) but this refers to a cell whose response peaks at about $2\,\mu$ and Felgett has cast some doubt on the reliability of this value. Fellgett (1949) quotes a value of W_m (reduced to an area of 0·1 cm^2) of 4×10^{-13} watt.

Methods of preparation of PbTe photo-conductive layers by various authors have already been discussed in § 2.6. A number of PbTe cells made at T.R.E. have been supplied to various research workers in high resolution infra-red spectroscopy. No actual figures for ultimate sensitivity have been given but from the published spectra (e.g. Thompson

Fig. 16

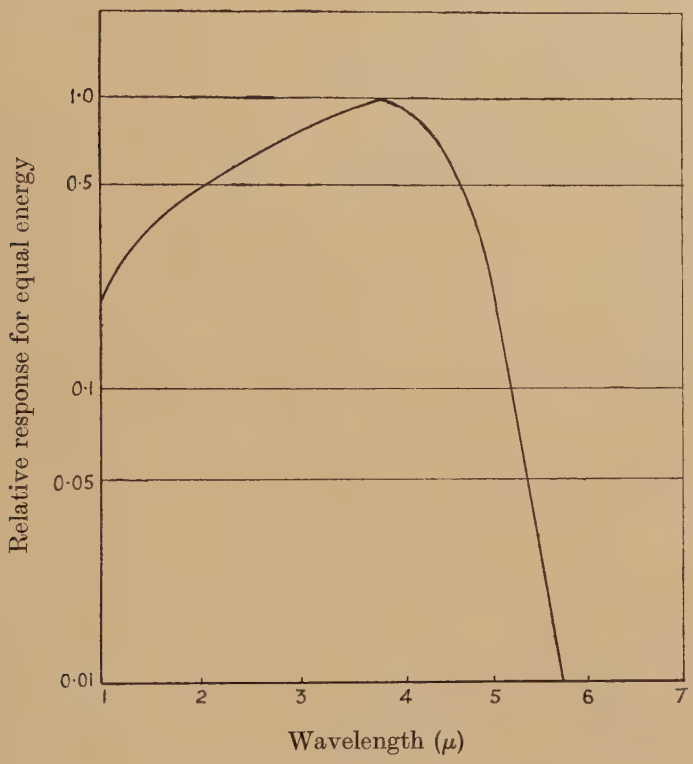


Spectral response of PbS cell at room temperature.

and Williams 1952, Boyd and Thompson 1952) obtained with these cells it is clear that they must have minimum detectable energies of the order of 10^{-13} watt for a 1 c/s bandwidth. In a recent review of PbTe and PbSe cells Moss (1953) quotes Thompson as obtaining a resolution of 0·15 cm^{-1} with a 4-in. grating at $3\,\mu$ using a T.R.E. PbTe cell. This would indicate an ultimate sensitivity of about 500 times better than a thermocouple. In addition, these cells may have a very rapid response—of the order of a few microseconds and so may be used to record spectra at a much higher rate than can be obtained with a thermocouple.

The T.R.E. cells are made by evaporating a layer of PbTe in a low pressure of oxygen and several variants of the method are described by Moss (1953). The layers are formed in a dewar-type glass cell. The incident radiation passes to the sensitive layer through a sapphire window fixed to the glass cell by means of a graded seal. These cells shew only very slight sensitivity at room temperature and are always operated at the temperature of liquid air. The spectral response of a modern PbTe cell (T.R.E.) at liquid air temperature is shewn in fig. 7. The $\lambda_{1/2}$ value is about $4{\cdot}75\,\mu$ and the sensitivity falls to 1% of its peak value at about $5{\cdot}75\,\mu$.

Fig. 17



Spectral response of PbTe cell at 90°K.

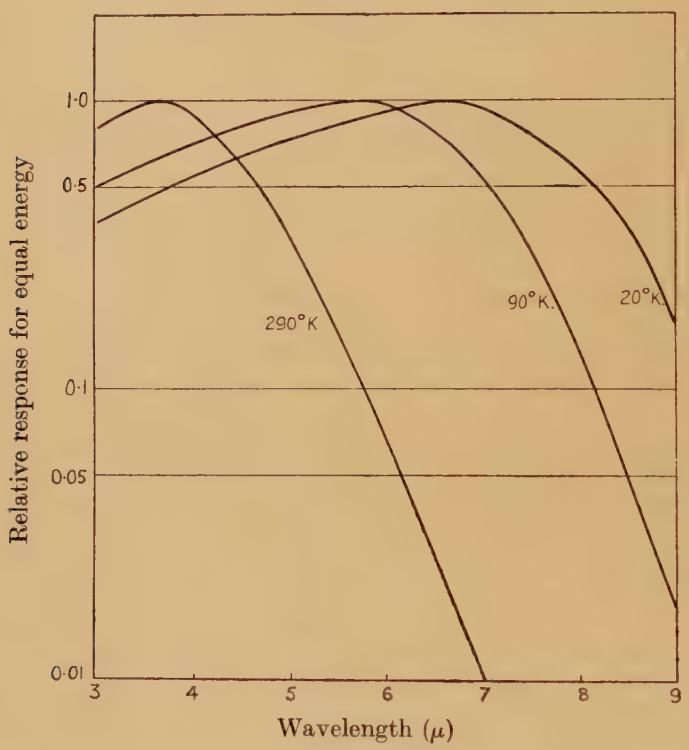
7.3. *PbSe Photo-conductive Infra-red cells*

The discovery by Gibson, Lawson, and Moss (1951) that PbSe cells could be made with a long-wave cut-off well beyond that found for PbTe led to an unexpected advance in infra-red detectors and opened up new regions of the spectrum to the use of photo-electric methods. Various authors had previously described PbSe cells but these did not exceed PbTe cells in long-wave response and seemed inferior as regards ultimate sensitivity except at room temperature, when PbSe has considerably

greater sensitivity. Methods of preparation used by various workers have already been described in § 2.6.

In his recent review Moss (1953) gives further information on PbSe cells made at T.R.E. which are sensitive to beyond 9 μ at liquid air temperature. The method of preparation is very similar to that used for PbTe cells. In this case windows of periclase (MgO) are used, as sapphire cuts off at too short a wavelength to enable full use to be made of the long-wave response of the cells. Spectral response curves for a typical T.R.E. cell (reduced to the same value at the maximum) are shewn in fig. 18 for three

Fig. 18



Spectral response of PbSe cell for 290°K, 90°K and 20°K.

temperatures. At liquid air temperature the maximum sensitivity of PbSe cells made so far is not so high as for the best PbTe cells but is considerably better than for a thermocouple, and they are very much faster, having a time-constant of the order of 10 μs. Using such a cell Roberts and Young (1953) have observed the centre of the water-vapour band at 6 μ at a rate of 50 spectra per second. They have also obtained photo-electric recordings of the complete water-vapour band from 5 μ to 7 μ at normal recording speed with a resolution of the order of 1 cm^{-1}. This, however, was set by the spectrometer rather than the cell.

It is interesting to note that PbSe cells may be made to operate at room temperature like PbS cells. So far, however, their room temperature sensitivity is considerably lower. Starkiewicz (1948) quotes a value for W_m of 2×10^{-9} watt for 1 c/s bandwidth. This is a fairly old figure and it is reasonable to suppose that, as for all other types of cells, more recent work has led to greater sensitivity. No more recent figures have, however, been published for uncooled PbSe cells but there is no doubt that a large gain in sensitivity is obtained on cooling to liquid air temperature. No further gain in ultimate sensitivity is obtained on cooling to liquid hydrogen temperature but the long-wave response is extended to give appreciable sensitivity out to nearly $10\,\mu$. The $\lambda_{1/2}$ values for 290°K, 90°K, 20°K are respectively $4{\cdot}7\,\mu$, $7{\cdot}1\,\mu$, $8{\cdot}2\,\mu$ with sensitivity down to 1% at about $7\,\mu$, $9{\cdot}3\,\mu$, $10{\cdot}2\,\mu$ (extrapolated).

7.4. *Limit to Sensitivity of Photo-conductive Cells*

If the electrical noise generated by the sensitive layer were negligible, the ultimate limit to the sensitivity of a photo-conductive cell would be set by the fluctuations in the radiation incident on the cell. This has been discussed by Fellgett (1949) and by Moss (1950 d) who have calculated the ultimate limits for a number of cells. They have shewn that, provided the quantum energy corresponding to the long-wave 'limit' is less than about 5 kT, which holds in practice for the cells under discussion, the effect of radiation fluctuations may simply be calculated from the fluctuation ΔN of the number of 'effective' quanta N_e incident on the layer per second using the simple 'classical' formula for the mean square value

$$\overline{\Delta N^2}=N_e \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

N_e is given by the expression $q_m n_e$ where

$$n_e=\frac{1}{h\nu_m}\int_0^\infty s(\nu)W(\nu)\,d\nu \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

and q_m is the quantum efficiency at the frequency ν_m corresponding to maximum sensitivity. $s(\nu)$ is the 'equal energy' spectral response of the cell normalized to unity at the frequency ν_m and $W(\nu)\,d\nu$ is the radiant energy falling on the layer in the frequency interval $\nu-\nu+d\nu$. $W(\nu)$ is usually simply given by the Planck distribution for room temperature radiation incident on the layer. One or two sides of the layer will be effective according to whether it is cooled or not. Now suppose a number N_s of effective quanta absorbed by the layer produce a signal voltage $\propto V_s$, the voltage fluctuation ΔV corresponding to the fluctuation ΔN will be given by

$$\overline{\Delta V^2}=\alpha^2\overline{\Delta N^2}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

If we refer to a bandwidth Δf we have

$$\overline{\Delta V^2}=2\alpha^2\overline{\Delta N^2}\Delta f \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

(cf. shot effect).

The signal voltage V_m when energy W_m of frequency ν_m falls on the layer will be given by

$$V_m = \alpha q_m W_m / h\nu_m. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

The value of W_m conventionally taken as the minimum detectable energy will be obtained when $V_m^2 = \Delta\overline{V^2}$, it will therefore be given by the expression

$$W_m = \left(\frac{2\Delta f}{q_m}\right)^{1/2} h\nu_m n_e^{1/2}. \quad . \quad . \quad . \quad . \quad . \quad (13)$$

Fellgett (1949) and Moss (1950 d) have evaluated W_m in this way, assuming $q_m = 1$, for a number of cells. For example, Fellgett finds $W_m = 3{\cdot}5 \times 10^{-13}$ watt for the particular PbS cell referred to in § 7.1 whereas the measured value (reduced to an area of 0·1 cm²) is 8×10^{-13} watt. This value is only 2·3 times greater than the ultimate limit set by radiation fluctuations. For a cooled PbTe cell he finds an even smaller factor, namely 1·9. It is thus clear that the best photo-conductive cells have sensitivities very close to the ultimate limit and moreover that the value of q_m, the quantum efficiency at the maximum of the response curve cannot be very different from unity. For most good modern cells the response per quantum is practically constant over a wide range of wavelengths, as we have seen, and it is very tempting to suppose that indeed $q_m = 1$ over this range. The limit in actual cells is normally set by current noise which has a $1/f$ type of spectrum whereas the spectrum of the radiation fluctuations would be expected to be uniform, at least at frequencies small compared with $1/\tau$ where τ is the time-constant of the cell.

Working from reasoning as above Simpson (1948) and Watts (1949), have suggested that if room temperature radiation were screened entirely from the photo-conductive layer when cooled to say 90°K the ultimate sensitivity should be increased, since the radiation fluctuations would be reduced. They have also shewn experimentally that this increase does take place. For PbTe cells, and also for certain PbS cells, screening off the room temperature radiation may increase the resistance of the layer by a factor of 100 or so, and also leads to a considerable reduction in minimum detectable energy. A marked increase in responsivity has also been observed by Simpson and Sutherland (1951). The reduction of minimum detectable energy is, however, not entirely due to the reduction of radiation fluctuations. The increase in resistance decreases the current for a given exciting voltage and this in turn reduces the current noise. The effect should therefore be observed even in cells limited by current noise. Unfortunately, full advantage cannot be taken of the considerable decrease in minimum detectable energy, amounting to about ten times in some cases, with cells as generally used in spectroscopy. A fairly wide aperture is normally required to permit entrance of the observed radiation and this exposes the layer to some room temperature radiation. However, if a fairly narrow aperture can be tolerated considerable gain may be obtained. Under these conditions

PbTe cells have been used shewing a minimum detectable energy less than that given theoretically from radiation fluctuations alone assuming the sensitive layer unscreened from room temperature radiation on one side.

Acknowledgments

I am indebted to many of my colleagues at T.R.E. for valuable discussions on the various topics dealt with in this review. Figures 3, 5, 7, 8, 9, 10, 11 are reproduced from the *Proceedings of the Physical Society* and figs. 1, 2, 4, 6 are based on figures appearing in *Semi-conducting Materials* (Butterworth). I am most grateful to the authors of the papers from which these figures are taken and to the publishers for permission to use them. Acknowledgment is made to the Chief Scientist, Ministry of Supply, for permission to publish this paper. Crown copyright reserved, reproduced by permission of the Controller, H.M. Stationery Office.

References

Arianova, I. I., and Sokolskaya, I. L., 1951, *Z. Tech. Fiz.*, **21,** 713.
Avery, D. G., 1951, *Proc. Phys. Soc.* B, **64,** 1087 ; 1952, *Ibid.*, **65,** 425 ; 1953, *Ibid.*, **66,** 134.
Banbury, P. C., 1952, *Proc. Phys. Soc.* B, **65,** 236.
Banbury, P. C., Gebbie, H. A., and Hogarth, C. A., 1951, *Semi-conducting Materials*, ed., H. K. Henisch (Butterworth), p. 78.
Banbury, P. C., and Henisch, H. K., 1950, *Proc. Phys. Soc.* B, **63,** 540.
Bauer, K., 1940, *Ann. der Phys.*, **38,** 84.
Becker, M., and Fan, H. Y., 1950, *Phys. Rev.*, **78,** 301.
Bell, D. G., Hum, D. M., Pincherle, L., Sciama, D. W., and Woodward, P. M., 1953, *Proc. Roy. Soc.* A, **217,** 71.
Blackwell, D. E., Simpson, O., and Sutherland, G. G. B. M., 1947, *Nature, Lond.*, **160,** 793.
Bottom, V. E., 1948, *Phys. Rev.*, **74,** 1218 ; 1949, *Ibid.*, **75,** 1310.
Boyd, D. R. J., and Thompson, H. W., 1952, *Trans. Far. Soc.*, **48,** 493.
Braithwaite, J. G. N., 1951, *Proc. Phys. Soc.* B, **64,** 274.
Busch, G., Wieland, J., and Zoller, H., 1951, *Semi-conducting Materials*, ed., H. K. Henisch (Butterworth), p. 49.
Cashman, R. J., 1946, *Jour. Opt. Soc. Am.*, **36,** 356.
Chasmar, R. P., 1948, *Nature, Lond.*, **161,** 281.
Chasmar, R. P., and Putley, E. H., 1951, *Semi-conducting Materials*, ed., H. K. Henisch (Butterworth), p. 208.
Chasmar, R. P., and Gibson, A. F., 1951, *Proc. Phys. Soc.* B, **64,** 595.
Clark, M. A., and Cashman, R. J., 1952, *Phys. Rev.*, **85,** 1043.
Darby, J., Hatton, J., and Rollin, B. V., 1950, *Proc. Phys. Soc.* A, **63,** 1181.
Dunaev, C. A., and Maslakovitz, J. P., 1947, *Jour. Exp. Theor. Phys.*, **10,** 901.
Eastman Kodak, Co., 1952, *Data sheet for Elektron detectors.*
Ehrenberg, W., and Hirsch, J., 1951, *Proc. Phys. Soc.* B, **64,** 700.
Eisenmann, L., 1940, *Ann. der Phys.*, **38,** 121.
Elliott, A., 1947, *Electronics, and their Application in Science and Industry* (Pilot Press), p. 97.
Ewald, A. W., 1951, *Phys. Rev.*, **81,** 607.
Fellgett, P., 1949, *Jour. Opt. Soc. Am.*, **39,** 970.

Fukuroi, T., Tanuma, S., and Tobisawa, S., 1949 a, *Sci. Rep. Res. Inst. Tôhuku Univ.*, **A1,** 365 ; 1949 b, *Ibid.*, 375 ; 1950 a, *Ibid.*, **A2,** 233 ; 1950 b, *Ibid.*, 239.

Gebbie, H. A., Banbury, P. C., and Hogarth, C. A., 1950, *Proc. Phys. Soc.* B, **63,** 371.

Gibson, A. F., 1949, *Nature, Lond.*, **163,** 121 ; 1950, *Proc. Phys. Soc.* B, **63,** 756 ; 1951, *Ibid.*, **64,** 603 ; 1952 a, *Ibid.*, **65,** 196 ; 1952 b, *Ibid.*, **65,** 214 ; 1952 c, *Ibid.*, **65,** 378.

Gibson, A. F., Lawson, W. D., and Moss, T. S., 1951, *Proc. Phys. Soc.* A, **64,** 1054.

Gibson, A. F., and Moss, T. S., 1950, *Proc. Phys. Soc.* A, **63,** 176.

Goldberg, L., 1950, *Reports on Progress in Physics*, **13,** 24.

Granville, J. W., and Hogarth, C. A., 1951, *Proc. Phys. Soc.* B, **64,** 488.

Halvosen, K. G., 1951, *Thesis*, Northwestern University.

Hatton, J., Rollin, B. V., and Seymour, E. F. W., 1951, *Proc. Phys. Soc.* A, **64,** 667.

Henisch, H. K., and Granville, J. W., 1951, *Semi-conducting Materials*, ed., H. K. Henisch (Butterworth), p. 87.

Hintenberger, H., 1942, *Zeit. f. Phys.*, **119,** 1.

Hogarth, C. A., 1953, *Proc. Phys. Soc.* B, **66,** 216.

Hogarth, C. A., and Granville, J. W., 1951, *Proc. Phys. Soc.* B, **64,** 992.

Hudson, R. P., 1951, *Proc. Phys. Soc.* A, **64,** 751.

Humphrey, J. N., Lummis, F. L., and Scanlon, W. W., 1952, *Phys. Rev.*, **86,** 660.

James, H. M., 1949, *Science*, **110,** 254.

Johnson, V. A., 1948, *Phys. Rev.*, **74,** 1255.

Kendall, J. T., 1950, *Proc. Phys. Soc.* B, **63,** 821.

Kicinski, F., 1948, *Chemistry and Industry*, **17,** 54.

Kolomiets, B. T., 1948, *J. Tech. Phys. U.S.S.R.*, **18,** 1456.

Lawson, W. D., 1951, *Jour. Appl. Phys.*, **22,** 1444 ; 1952, *Ibid.*, **23,** 495.

Legrand, K., 1948, *Zeit. f. Phys.*, **124,** 219.

Loferski, J. J., and Miller, P. H., 1951, *Phys. Rev.*, **83,** 876.

Lothrop, J. W., 1949, *Thesis*, Northwestern University.

Milner, C. J., and Watts, B. N., 1949, *Nature, Lond.*, **163,** 322 ; 1952, *Research*, **5,** 267.

Mitchell, E. W. J., and Sillars, R. W., 1949, *Proc. Phys. Soc.* B, **62,** 509.

Moss, T. S., 1947, *Nature, Lond.*, **159,** 76 ; 1948, *Ibid.*, **161,** 776 ; 1949 a, *Proc. Phys. Soc.* A, **62,** 264 ; 1949 b, *Ibid.*, B, **62,** 741 ; 1950 a, *Ibid.*, **63,** 167 ; 1950 b, *Ibid.*, **63,** 982 ; 1950 c, *Phys. Rev.*, **79,** 1011 ; 1950 d, *Jour. Opt. Soc. Am.*, **40,** 603 ; 1951, *Proc. Phys. Soc.* A, **64,** 590 ; 1952 a, *Ibid.*, B, **65,** 147 ; 1952 b, *Photoconductivity in the elements* (Butterworth) ; 1953, *Research*, **6,** 258.

Moss, T. S., and Chasmar, R. P., 1948, *Nature, Lond.*, **161,** 244.

Mott, N. F., and Gurney, R. W., 1940, *Electronic Processes in Ionic Crystals* (Oxford), p. 158.

National Bureau of Standards, 1948, Table 27–4.

Paul, W., and Jones, R. V., 1953, *Proc. Phys. Soc.* B, **66,** 194.

Paul, W., Jones, D. A., and Jones, R. V., 1951, *Proc. Phys. Soc.* B, **64,** 528.

Pincherle, L., 1951, *Proc. Phys. Soc.* A, **64,** 648.

Putley, E. H., 1952 a, *Proc. Phys. Soc.* B, **65,** 388 ; 1952 b, *Ibid.*, **65,** 736 ; 1952 c, *Ibid.*, **65,** 992.

Putley, E. H., and Arthur, J. B., 1951, *Proc. Phys. Soc.* B, **64,** 616.

Reimann, A. L., and Sullivan, J. V., 1952, *Proc. Phys. Soc.* B, **65,** 480.

Rittner, E. S., 1950, *Science*, **111,** 685.

Rittner, E. S., and Grace, F., 1952, *Phys. Rev.*, **86,** 615.

Roberts, V., and Young, A. S., 1953, *Jour. Sci. Inst.*, **30,** 199.
Rollin, B. V., and Simmons, E. L., 1952, *Proc. Phys. Soc.* B, **65,** 995 ; 1953, *Ibid.*, **66,** 162.
Russell, B. R., and Wahlig, C., 1950, *Rev. Sci. Inst.*, **21,** 1028.
Scanlon, W. W., Petritz, R. L., and Lummis, F. L., 1952, *Phys. Rev.*, **86,** 659.
Schwarz, E., 1948, *Nature, Lond.*, **162,** 614 ; 1949, *Proc. Phys. Soc.* A, **62,** 530 ; 1950, *Ibid.*, B, **63,** 624 ; 1951, *Ibid.*, **64,** 821.
Shockley, W., 1951, *Electrons and Holes in Semi-conductors* (Van Nostrand).
Simpson, O., 1947, *Nature, Lond.*, **160,** 791 ; 1948, *Proc. Phys. Soc.*, **61,** 486.
Simpson, O., Sutherland, G. G. B. M., and Blackwell, D. E., 1948, *Nature, Lond.*, **161,** 281.
Simpson, O., and Sutherland, G. G. B. M., 1951, *Phil. Trans. Roy. Soc.* A, **243,** 547 ; 1952, *Science*, **115,** 1.
Smith, R. A., 1950, *Science*, **112,** 71 ; 1951, *Semi-conducting Materials*, ed., H. K. Henisch (Butterworth), p. 198.
Sosnowski, L., 1947, *Phys. Rev.*, **72,** 641.
Sosnowski, L., Soole, B. W., and Starkiewicz, J., 1947, *Nature, Lond.*, **160,** 471.
Sosnowski, L., Starkiewicz, J., and Simpson, O., 1947, *Nature, Lond.*, **159,** 818.
Starkiewicz, J., 1948, *Jour. Opt. Soc. Am.*, **38,** 481.
Starkiewicz, J., Sosnowski, L., and Simpson, O., 1946, *Nature, Lond.*, **158,** 28.
Strong, J., 1951, *Physics Today*, **4,** pt. 4, 14.
Sutherland, G. G. B. M., and Lee, E., 1948, *Reports on Progress in Physics*, **11,** 144.
Thompson, H. W., and Williams, R. L., 1952, *Trans. Far. Soc.*, **48,** 502.
Watts, B. N., 1949, *Proc. Phys. Soc.* A, **62,** 456.
Williams, V. J., 1948, *Rev. Sci. Inst.*, **19,** 135.
Wilman, H., 1948, *Proc. Phys. Soc.* **60,** 117.
Wyrick, R., and Levinstein, H., 1950, *Phys. Rev.*, **78,** 304.
Zhuze, V. P., Mochan, I. V., and Ruirkins, S. M., 1948, *J. Tech. Phys. U.S.S.R.*, **18,** 1494.

We beg to acknowledge the source of the following text-figures :—

Figure 3 from Gibson, 1950, *Proc. Phys. Soc.* B, **63,** 759.
Figure 5 from Moss, 1949, *Proc. Phys. Soc.* B, **62,** 743.
Figures 7–11 from Gibson, 1952, *Proc. Phys. Soc.* B, **65,** 381–384.

Thermodynamic and Kinetic Properties of Glasses

By R. O. Davies and G. O. Jones
Department of Physics, Queen Mary College, University of London

CONTENTS

§ 1. Introduction

The most important advances in the study of the nature of the glassy state were the results of measurements of the specific heats of liquids and super-cooled liquids, carried out many years ago with the immediate aim of determining the range of validity of thermodynamic laws. An account of these early developments provides a very convenient approach to our main subject.

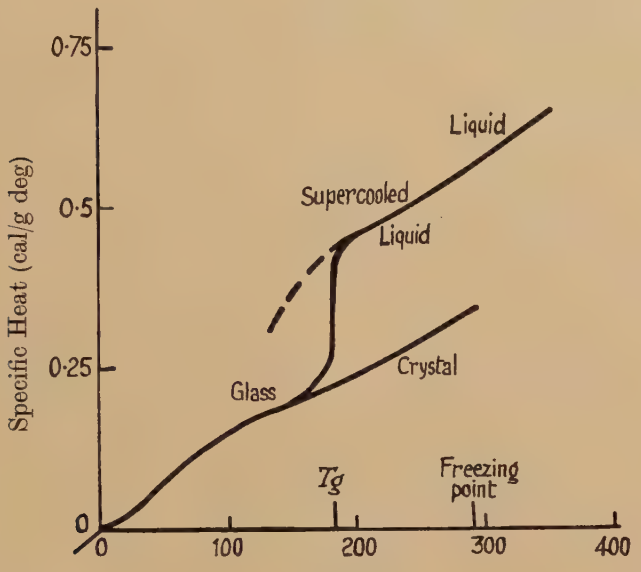
1.1. *Thermal Properties of Amorphous Substances*

The scientific importance of the glassy state was first recognized by Lewis and Gibson in 1920 when they predicted that super-cooled liquids would be exceptions to the Nernst Heat Theorem. Their argument was actually misleading in that they regarded liquids as associated compounds which should retain an entropy of mixing down to the lowest temperatures. However, their prediction was correct and was soon confirmed by Wietzel (1921) for fused silica and by Gibson and Giauque (1923) for glycerol. These authors continued their measurements down only to liquid air temperatures so that there remained the possibility of a specific heat 'anomaly' at still lower temperatures. This possibility was excluded by Simon and Lange (1926) who measured the heat capacity of glycerol down to 10°K. Their results, which are typical of the behaviour of a wide range of amorphous substances, are illustrated in

fig. 1. As will be seen, the specific heat of supercooled liquid glycerol shows a rapid decrease (from about 0·45 to 0·23 cal/g deg) as the temperature is taken down through a fairly narrow range in the neighbourhood of 180°K. Below this temperature the specific heat of the supercooled liquid is practically the same as that of crystalline glycerol.

Calculation shows that at the lowest temperatures the entropy of the supercooled amorphous material exceeds that of the crystal by 4·6 cal/mole deg, so that the system appears to violate the Heat Theorem. However, it was pointed out by Simon (1930) that the discrepancy would disappear if the specific heat curve of the amorphous substance were to follow at low temperatures a reasonable extrapolation of the curve taken

Fig. 1



Specific Heat of Glycerol (after Simon and Lange.)

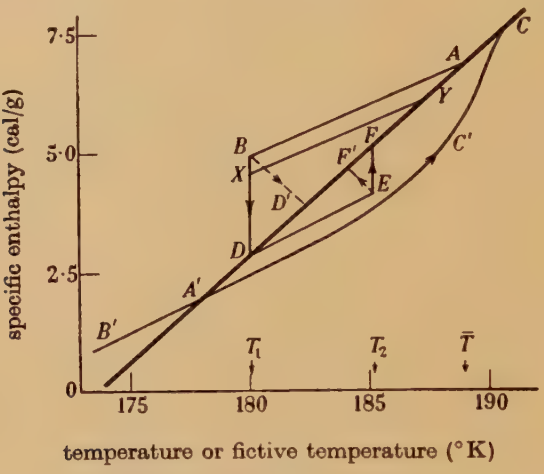
at higher temperatures, without showing the rather sudden characteristic drop (a suitable extrapolated curve is shown as a broken line in fig. 1). It was later shown by Oblad and Newton (1937) that, if measurements of specific heat were made very much more slowly, the beginnings of such a curve could actually be observed and that the fall in specific heat was moved to a somewhat lower temperature.

The temperature region over which the drop in specific heat occurs cannot therefore be related exactly to any thermodynamic property of the material, but depends on the experimental methods employed, particularly on the ' time-scale ' of the experiment. It depends also on the second independent thermodynamic variable (such as the pressure)

which may be used to specify the state. Experiments performed under atmospheric pressure with a time-scale of the order of minutes will show these relaxation effects, or ' transformation ' phenomena, over a range of temperature *roughly* defined by a ' transformation temperature ', T_g.

Above T_g, the substance is said to be a liquid (or a supercooled liquid if the temperature is below the freezing point), and below T_g it is defined to be a *glass*. The glass is thermodynamically unstable in a sense which differs from that in which the metastable supercooled liquid is unstable and it will, in principle, ultimately move irreversibly into one of a series of states which are continuous with states above T_g. This process is sometimes called ' stabilization ' ; it is strongly influenced by temperature, and it is one of the processes which play a part in the annealing of ordinary glass.

Fig. 2



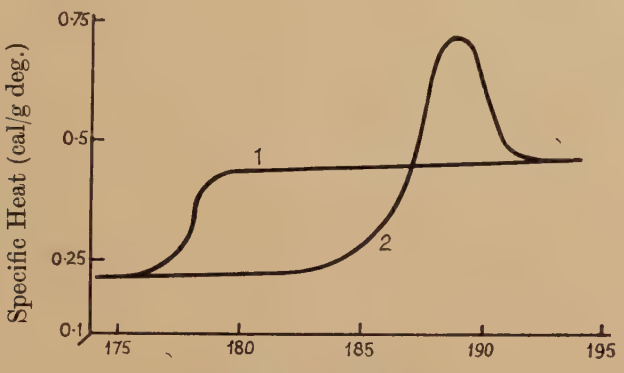
Glass Formation in Glycerol.

In order to describe more precisely the behaviour of amorphous substances near their transformation temperatures it is convenient to keep in mind the curves illustrated in fig. 2 and in fig. 3. In fig. 2 the enthalpy, H, of glycerol is plotted against temperature. (Any other extensive variable such as the volume or entropy would exhibit some generally similar dependence upon temperature.) The curve corresponding to the equilibrium supercooled liquid is labelled CD. According to the rate of cooling employed the glycerol will depart from its equilibrium behaviour at some point, such as A or A′, and proceed along AB or A′B′—the former path being followed with the more rapid cooling. If, after moving some distance along AB, the temperature is kept constant at T_1, the value of the property in question will move along BD at a rate which depends mainly on T_1, being more rapid for higher T_1. The stabilized glass at D can now be heated to state E, say (at temperature T_2) and

again kept at constant temperature. It will then return to equilibrium along EF, the relaxation process taking less time than at the lower temperature.

In fig. 3 are plotted the possible results of two continuous measurements of specific heat. The results of such experiments depend both on the rate at which the glass is initially cooled and the rate at which it is heated during the ' run '. For both the curves 1 and 2 the glass is assumed to have been cooled slowly along A′B′. Curve 1 corresponds to observations made during heating at the same slow rate and curve 2 to those conducted at a faster rate (which as a rate of cooling, would have *formed* the glass along AB). The heating curve for the latter is given by B′A′C′C on fig. 2 and shows the reason for the characteristic hump often obtained experimentally. (The form of this hump has been given a quantitative explanation by Haseda, Otsubo and Kanda (1950).)

Fig. 3



Apparent Specific Heat of Glycerol.
1—slow rate of heating.
2—fast rate of heating.

Tammann (1933) showed that the viscosity of any glass-forming substance at T_g was always about 10^{13} poise. Although we shall see later that it is really a *volume* viscosity which should be used as a determinant, the two quantities seem to be roughly of the same order of magnitude so that in practice Tammann's criterion is a reasonable guide. In fact any liquid cooled without crystallization to a temperature at which its viscosity reaches this value becomes a glass. Glasses are of very diverse physical natures and a representative list of substances with their transformation temperatures could include: ethyl alcohol (95°K), $Na_2SO_3 . 5H_2O$ (230°K), polystyrene (350°K), ordinary silica glasses (~800°K) and silica (~1500°K). High polymers usually have rather sharply defined transformation temperatures because the rate of change in relaxation rate with temperature is very large (Boyer and Spencer 1946, Gee 1947, Alfrey 1948).

While the main facts about the variation of properties with temperature near the transformation temperature are now well established, little is known about their variation with pressure. At a given temperature we should expect to be able to define a ' transformation pressure ', p_g, above which the liquid would behave as a glass, that is, show smaller specific heat, expansion coefficient and compressibility than would be expected by extrapolation of data obtained at lower pressures. This has been confirmed by the experiments of Tammann and Jellinghaus (1929) (on salicylin) and of Scott (1935) (on rubber). It may provide an explanation for the observed increase of seismic velocities in the central core of the earth (Elsasser 1950) provided we assume that, at the frequency of the wave transmitted, p_g is the pressure at the boundary of the core.

The behaviour so far described is characteristic of all substances which can be obtained as glasses. In all these substances there is a temperature range in which relaxation effects can be observed, and in which discontinuities are shown in the specific heat and coefficient of expansion as ordinarily determined. However, there are many other systems which are thermodynamically similar. A simple example is provided by a chemically reacting system in which the rate of reaction varies rapidly with temperature. Above a certain temperature T_g, say, the reaction can proceed, while below T_g it is inhibited and the system corresponds to a glassy state. For this reason any purely thermodynamic discussion of these phenomena in glasses will be applicable to a much wider class of systems, namely, to any system for which non-equilibrium states are continuous with equilibrium states.

1.2. *General Explanation*

The central problem is to explain in molecular terms the way in which the glass differs from the liquid and the nature of the change from the glass to the equilibrium liquid. In view of the complexity of glass-forming substances we cannot hope for a detailed microscopic theory. However, an important step towards understanding the behaviour of glasses was made by Simon in 1930 when he used ideas drawn from statistical mechanics to give a qualitative explanation of the existence of residual entropy at absolute zero. It will be useful to discuss this development in some detail since Simon's ideas have often been overlooked (see for instance : Winter 1946, Buchdahl and Nielsen 1950).

Experimental evidence from studies on x-ray diffraction (Warren 1940) suggests that, for an amorphous material, it would be reasonable to associate with each state a ' degree of order ' descriptive of its geometrical state of order. We may suppose that it is represented by some parameter z, where z is a function of temperature. A variation in z requires, in general, that energy be added to or taken from the system on account of the change in potential energy due to changing configuration.

Simon pointed out that as a glass is cooled through its transformation temperature the molecular diffusion which is necessary to effect the appropriate change in configuration is increasingly inhibited and finally becomes practically impossible. Thus the value of z will become fixed somewhere near the transformation temperature and that part of the specific heat corresponding to changes in potential energy will be eliminated below this temperature. The ' configurational ' contribution to any other property will similarly disappear. At the same time the system ceases to be in true internal thermodynamic equilibrium. If one waits for a less or greater time (depending on the temperature) the system will move to its equilibrium state with a different value of z. In this irreversible process heat will be liberated so that, in accordance with the unattainability principle, it is not possible to use this change for the production of cold. (In order to do this the system would have to change from a stabilized liquid into one of the series of unstabilized glasses ; such a change would violate the Second Law.)

To draw attention to the fact that a glass cannot be treated as a thermodynamic system if we are considering processes requiring a change of molecular arrangement, Simon suggests that the vapour pressure of a glass will almost certainly be indeterminate and change with time. Thus, according to the conditions of the experiment, one might expect that a layer of molecules deposited on the surface could either be amorphous or show a micro-crystalline structure. In any case it is unlikely to have the same structure as the main body of the glass. There is however, no objection to applying the Second Law to any process in which the configuration remains fixed, as indeed it does in all processes occurring at temperatures well below T_g.

It would obviously be desirable to know more about the physical significance of the ordering parameter z. Unfortunately it is at present possible only to explain in the most general terms what are thought to be the physical changes occurring when the temperature is varied and the structure is allowed to reach equilibrium. We might expect that in equilibrium the configuration becomes more ordered as the temperature is decreased, and this is confirmed to some extent by the results of x-ray diffraction studies on liquids. These are usually expressed in the form of curves of the ' radial distribution function ' of atoms plotted against their distance from an arbitrarily chosen central atom. For an ideal crystal such a curve would consist of a number of lines parallel to the axis along with the radial distribution function is measured, while the curve for an ideal gas consists of a parabola having its axis collinear with the same line. A hump in any curve above the course of this parabola indicates a more than random probability of finding a second atom within a given range of distances from a central atom, and the typical curve for a liquid consists of a small number of humps and troughs diminishing in size as the distance from the central atom increases. It thus indicates the existence of a

certain degree of ' structure ' in the liquid, persisting over a small number of inter-atomic distances, that is, of some ' short-range order '. The results obtained, mainly by Gingrich and co-workers (Gingrich 1943), on ordinary liquids such as the liquid alkali metals always show the same kind of variation with temperature. At lower temperatures the humps become narrower but higher, and the area below a given hump decreases. These changes indicate that there is an increase in order, or an increase in the degree of ' structure ', with falling temperature, since they reflect a reduction in the scatter of inter-atomic distances. They also indicate that the average co-ordination decreases, that is, the structure becomes more ' open ' at lower temperatures. These variations in equilibrium structure with temperature are in fact those predicted by Bernal's theory of liquids (1937), in which the first attempt was made to explain statistically the existence of a configurational specific heat. It would seem that a suitable ordering parameter (z) for these simple liquids might be some functional of the radial distribution function. Unfortunately, the situation is less simple for the more complex liquids which can be obtained as glasses. It is known that the properties of these liquids are determined largely by the directional nature of their main bonds, and the radial distribution function as experimentally determined gives no direct information about the relative orientations of the atoms or molecules concerned. We thus suppose that a suitable quantity z for these liquids would have to contain more information than given by the radial distribution function.

Evidence about the variation in structure with temperature for glass-forming liquids can of course be obtained by the use of x-ray methods, though in fact very little work has been done on this problem. The most immediately relevant results we can quote are those of Bernal and Fowler (1933) on water. In many respects water is a typical glass-forming substance (see Staronka 1939, Douglas and Isard 1951, Pryde and Jones 1952) even though it is difficult to obtain samples of water in the vitreous state. The progressive change again appears to have the same general sense and, as interpreted by Bernal and Fowler, consists of a change from a structure approximating to a close-packed arrangement at the highest temperatures to one approximating to a tetrahedral and quartz-like arrangement at lower temperatures.

We are now able to assess more clearly the role of viscosity. The adjustments which occur during variations in the configuration must involve small atomic or molecular rotations and migrations. Since approximately similar adjustments must occur during viscous flow it is to be expected that the viscosity enters in determining the temperature (T_g) below which the required configurational adjustments take very long times. It is clear that the adjustments are on the scale of atomic dimensions and it is indeed possible to obtain samples of glass which are in equilibrium from the point of view of our discussion but which contain visible defects or holes. Heating to temperatures much higher than T_g is necessary before the large-scale adjustments necessary to eliminate these can occur.

Finally, we should mention that the correctness of the basic supposition of Simon's picture—that the configuration in a glass remains more or less frozen-in below T_g—has not yet been investigated directly by x-ray methods. The evidence for its truth, though indirect, is still convincing, as we hope will appear in the more detailed discussion which follows.

§ 2. Static Properties of Super-Cooled Liquids and Glasses

We now discuss from the standpoint already outlined some implications of the experimental data on properties such as the specific heat, expansivity and compressibility—which would normally be thought of as the thermodynamic properties—of glass-forming substances. Because we have to refer to the properties both of meta-stable phases and of phases not in internal equilibrium we avoid the term ' thermodynamic ' at this stage and substitute ' static '.

2.1. *Experimental Data*

The following table summarizes experimental data from many sources on the static properties of super-cooled liquids and glasses for a number of substances in the neighbourhood of their respective transformation temperatures. The figures are collected from the papers of Tammann and Jellinghaus (1929), Kauzmann (1948) and Davies and Jones (1953).

Table 1

	T_g (°K)	C_p' (cal/g)	ΔC_p (cal/g)	α' ($\times 10^4$/deg)	$\Delta\alpha$ ($\times 10^4$/deg)	κ' ($\times 10^{12}$ cm^2/dyne)	$\Delta\kappa$ ($\times 10^{12}$ cm^2/dyne)
Glycerol	180	0·25	0·21	2·4	2·4	—	—
Glucose	300	0·33	0·18	0·90	2·6	9·3	6·1
Boron Trioxide	470–530	0·30	0·14	0·5	5·6	—	—
Selenium	300	0·08	0·045	1·7	2·5	24·4	5·8
Polyisobutylene	190–200	0·27	0·09	0·5	5·5	—	—
Rubber	200–320	0·26	0·13	2·0	3·1	27	37
Polystyrene	350	0·32	0·069	2·3	2·0	23	75
Colophonium	300	0·27	0·13	3 ca.	3 ca.	25 ca.	10 ca.

Values of the specific heat (C_p'), expansivity (α') and compressibility (κ') in the glassy state (just below T_g) are tabulated, with corresponding values of ΔC_p (equal to $C_p - C_p'$, where C_p refers to the super-cooled liquid just above T_g), $\Delta\alpha$ and $\Delta\kappa$. As we shall show later, knowledge of the values of the incremental quantities named is necessary before an evaluation can be made of a thermodynamic theory of the relations between properties of the glass and super-cooled liquid. Unfortunately, for only a very small number of substances are present data even nearly adequate for this purpose. This is partly because of the existence of inherent experimental difficulties in the problem, but largely because most workers have misdirected their efforts because there has been no accepted thermodynamic theory to serve as a guide to experiment. The substances chosen for insertion in table 1 are those for which at least two of the

quantities ΔC_p, $\Delta\alpha$ and $\Delta\kappa$ are known. It will be seen that the common inorganic glasses are hardly represented ; this is because their transformation temperatures are so high that accurate measurements of C_p, and even rough measurements of κ, are very difficult.

A striking fact which emerges from these results is that the ratios $\Delta C_p/C_p'$, $\Delta\alpha/\alpha'$ are of the order unity. We can fairly conclude that the configurational contribution to the specific heats of these liquids—and perhaps of all liquids—must be very large, and of the same order as the vibrational contribution. Though estimates of the configurational specific heats of liquids have been made in other ways (see for example, Staveley, Hart and Tupman 1953, Pople 1953), they are open to doubt because of uncertainty about the vibrational and rotational contributions to the specific heat. In a glass-forming liquid, however, the configurational specific heat may be measured directly at a single temperature, and it is worth considering whether this possibility might be important in the development of a general theory of liquids. The outlook is unfortunately not promising. While many rough theories of the liquid state—such as those of Bernal (1937), Lennard–Jones and Devonshire (1937, 1938), Mott and Gurney (1939) and Frenkel (1946)—include the idea that the equilibrium configuration is a function of temperature, and more refined theories contain a potential energy of interaction from which the configurational specific heat can in principle be derived, experiment and theory do not in fact make contact. For the only liquids which can form glasses are those which possess directed or localized bonds, and such liquids have so far been out of reach of theoretical analysis.

The values of ΔC_p, $\Delta\alpha$ and $\Delta\kappa$ given in table 1 are all greater than zero. However, there are cases in which $\Delta\alpha$ is negative. Douglas and Isard (1951) have recently reported that α is negative for silica at temperatures near T_g, although their conclusion was not the result of direct measurement of α above T_g but of a reasonable interpretation of data obtained in measurements of α' at room temperature, using specimens rapidly chilled from different temperatures in the neighbourhood of T_g. Since α' for silica is positive, though very small, this implies that $\Delta\alpha$ is negative. Again, it is well known that the expansivity (α) of water in the range immediately below 4°c is negative, although its behaviour near T_g is unknown. It will be shown later that there are thermodynamic reasons why ΔC_p and $\Delta\kappa$ must always be positive, though $\Delta\alpha$ may be either positive or negative, and that this is also true for the quite different so-called ' lambda transitions ', where the sign of ΔC_p, etc., now refers to the direction of the ' lambda ' relative to a smooth curve interpolated between portions lying at higher and lower temperatures.

The relations between the values of these properties for given substances as glass and crystal have not been systematically studied and insufficient data exist to make the addition of extra columns in the above table worth while. As shown for the case of glycerol in fig. 1, values of C_p for the glass and crystal are close to each other at all temperatures

below T_g, and the same appears to be true for other substances (see White 1919, Borelius and Paulson 1946). This is not surprising if we realize that the non-configurational contribution to C_p must be mainly vibrational in origin. The fact that C_p for the glass slightly exceeds its value for the crystal is clearly due to the greater volume of the glass, as suggested originally by Simon. Again, the expansivity of a glass and its crystal often appear to be nearly equal at the same temperature, although where there are several possible crystalline forms (as in silica) no clear statement can be made. For silica, also, the value of α' for the glass well below T_g can be altered by varying the cooling schedule (Douglas and Isard 1951), so that the value of α' itself is somewhat indeterminate. It is not possible from existing data to derive any corresponding conclusions about the compressibility. We do not pursue these points because the thermodynamic treatment presented later does not cover relationships between the properties of glasses and their crystals—which must in general be determined by the specific differences in structure between glass (or liquid) and crystal for the substance in question.

2.2. *The Super-cooled Liquid as a Meta-stable Phase*

Although a super-cooled liquid is thermodynamically unstable relative to its crystal, we have referred to the states of a super-cooled liquid as ' equilibrium ' states. This is reasonable if we are discussing only the irreversible approach towards the super-cooled liquid from glassy states because the glass is itself unstable relative to the super-cooled liquid. It is however important to distinguish between the two kinds of instability. A super-cooled liquid is in fact in a state which is thermodynamically quite analogous to that of, say, diamond—a condition for which the term ' meta-stable ' is usually reserved. Although diamond at ordinary temperatures and pressures is thermodynamically unstable relative to graphite, the transformation from diamond to graphite does not occur, and diamond obeys the Nernst Heat Theorem. Again, a sample of a super-cooled liquid may be maintained almost indefinitely, in the absence of freezing nuclei, except within a fairly narrow range of temperatures. Jones and Simon (1949) illustrate the condition of such systems in terms of a model consisting of a single particle lying in a potential well separated from a lower well by a fairly high barrier. The condition of a glass would be represented by a particle stuck somewhere on the side of a well. In cases of the first kind all neighbouring positions of the particle are unstable and from them the particle tends to return to the original position. In cases of the second kind displacements towards the lower side are so much more probable than displacements towards the upper side that any finite displacements tends to take the particle nearer the centre of the well. A rough explanation of this in physical terms would be as follows : before a nucleus of graphite can be formed in diamond, or a nucleus of a crystal in a super-cooled liquid, a rather large local energy barrier must be

overcome. In a glass, however, there is always the possibility somewhere in the structure of an easily achieved adjustment leading to a slightly more favourable atomic distribution. The question is complicated by the need to consider configurational free energy rather than configurational energy only. Also, the phenomenon is ' co-operative ' in a more complete sense than in order–disorder transitions such as occur in ferromagnetics. In these the lattice at least is fixed. In glasses or liquids there is no fixed lattice arrangement but only a continuous range of possible micro-states.

There is no doubt about the reality of the equilibrium in the super-cooled liquid, since it has been shown by many workers (Lillie 1936, Davies and Jones 1953) that the equilibrium corresponding to a given temperature can be approached from both sides—as we have already stated in discussing fig. 2. The two ' sides ' of the equilibrium state are assumed to correspond respectively to smaller and larger values of z. Indeed, a curve of free energy against z at constant pressure and temperature would resemble exactly the contour of the simple model we have described above. The thermodynamic discussion given later rests on the assumption that a single complete equation of state such as $f(G, p, T, z)=0$ can in principle represent all glassy as well as ordinary super-cooled states of the liquid phase. The equation of state of the ' equilibrium ' super-cooled liquid is then found by applying the equilibrium condition $(\partial G/\partial z)_{p,T}=0$. Rather similar procedures are used in many approximate statistical theories covering other systems, although other additional independent variables (such as the volume) are often used, instead of an ordering variable. An interesting semi-empirical account of our problem in such terms is due to Borelius and Paulson (1946). Using the results of their measurements of the thermal properties of super-cooled and crystalline selenium, and assuming that the differences $H_{\text{liq}}-H_{\text{crys}}$ and $S_{\text{liq}}-S_{\text{crys}}$ are determined by the configuration of the super-cooled liquid (or glass) only, and not by the temperature (a reasonable assumption, as we have seen), it is possible to plot values of $G_{\text{liq}}-G_{\text{crys}}$ against $V_{\text{liq}}-V_{\text{crys}}$ at constant temperature. Such a curve would of course include points representing non-equilibrium states as well as equilibrium states. The curves obtained at different temperatures in fact all showed minima in the free energy difference, at values of the volume difference equal to those actually found for the equilibrium super-cooled liquid.

While Simon's explanation of the behaviour near T_g is undoubtedly correct, a more subtle question remains in connection with the thermodynamic status of the equilibrium super-cooled liquid. Thus it seems that it should be possible in principle to follow the curve of the equilibrium liquid—as shown dotted in fig. 1—down to absolute zero by waiting for a sufficient time. Such a system would be expected to obey the Nernst Heat Theorem (like diamond) and its entropy would be zero. The question arises as to what would be the structure of this liquid of zero entropy. If crystalline, this would imply the possibility of a continuous change from an amorphous to a crystalline state. Although this possibility cannot be

ruled out, experimental evidence obtained on the melting curve (Holland, Huggill and Jones 1951) and on the thermal properties of helium under pressure (Dugdale and Simon 1951) and the theoretical arguments of Bernal (1937) and of Landau and Lifshitz (1938), suggest that it is unlikely. The discussion has been given a new turn by Kauzmann (1948), who, by extrapolation to absolute zero of the curves obtained by plotting the values of certain extensive properties of the equilibrium liquid against temperature, concludes that they would suggest the existence at zero temperature of a liquid phase having *smaller* volume and entropy than the crystal. He disposes of this uncomfortable suggestion by means of an ingenious explanation which says, in effect, that any attempt to resolve the paradox is operationally meaningless. If we wish to discuss the properties of an equilibrium liquid state near absolute zero we must have some means of producing such a state—at least in principle. Now the barrier against crystallization in the super-cooled liquid can be thought of as divided into a barrier against nucleation and a barrier against flow. At high temperatures the former is more important. At low temperatures the barrier against flow—which must be closely related to the barrier limiting the irreversible approach to equilibrium from glassy states—dominates the expression for the rate of crystal growth. By considering the variation with temperature in the equilibrium structure of the *solid* which would be suggested by the model of Mott and Gurney (1939), he concludes that below a certain ' pseudo-critical ' temperature the equilibrium liquid could not be formed because crystallization would always occur first. (Frenkel (1935, 1937) has made similar suggestions in a somewhat different context.) Although there are certain objections to the procedures used by Kauzmann in his extrapolation of data, we may note that the method which he uses to resolve his paradox, if justified, disposes also of our original question. The problem of imagining a structure for an ordered equilibrium liquid at absolute zero now becomes a metaphysical rather than a physical one.

Finally, we mention briefly an attempt by Pauling and Tolman (1925) to explain quantitatively by statistical-mechanical methods the magnitude of the residual entropy of a glass at absolute zero. There are a number of weaknesses in their treatment but we refer only to the outstanding objection from our point of view. Because their starting-point is an attempt to estimate the number of positions which can be occupied by a *single* particle in the structure they ignore entirely the existence of cooperative effects and interactions. At best their calculation gives an estimate of the entropy of a liquid whose molecules have an arbitrarily selected number of equivalent positions. It does not make contact with the real problem, which would be to estimate the configurational entropy of the liquid at the temperature at which its viscosity was about 10^{13} poise.

2.3. *An Examination of some Past and Present Misconceptions*

Even a cursory study of publications dealing with the properties of the ordinary inorganic glasses shows that the phenomena occurring near the transformation temperature are still not everywhere understood.

It is often stated, for instance, that there is a ' sharp contraction ' or ' sharp increase in viscosity ' near T_g—suggestions which are of course quite false. The discontinuity is shown when the *expansivity*, not the volume, is plotted against temperature. More commonly found are suggestions that some kind of ' polymerization ' sets in below T_g. A more correct statement would be that a process analogous to a polymerization occurs continuously as the temperature is lowered *above* T_g, but that below T_g it cannot continue. It is the failure of the material to continue its polymerization (or ordering) which is responsible for the drop in the values of derivative properties as the temperature is lowered through T_g.

The latter misconception is somewhat similar in basis to another, which is implied by the frequent use of the terms ' second-order transition ' or ' apparent second-order transition ' in discussions of the glassy behaviour of high polymers and rubber. The terms are used presumably because the discontinuities in specific heat shown by glassy systems are somewhat similar to the sharp discontinuities envisaged by Ehrenfest in defining *ideal* second-order transitions, of which only one example appears to be known in nature (the behaviour of superconductors at their normal transition temperatures in zero magnetic field). There is added danger of confusion because ' lambda ' transitions are often referred to as second-order transitions, and because non-equilibrium effects exactly analogous to those with which we are concerned can occur in systems which show lambda transitions. It will be worth while to set out clearly where the differences and similarities lie.

All lambda transitions are basically equilibrium transitions. We consider a well-known example, the lambda transition shown by certain substitutional binary alloys of simple atomic compositions (such as 1 : 1 or 1 : 3). Some such alloys show a continuous change from order to disorder, as the temperature is raised, in respect of the distribution of the two kinds of atoms over the fixed lattice sites. At any temperature there is an equilibrium degree of order. However, because of cooperative effects the disordering occurs over a fairly narrow range of temperature and a sharp lambda is shown in the curve of specific heat against temperature. Even in these systems it is possible by rapid cooling to a given temperature to cause to be frozen-in a greater degree of disorder than corresponds to that temperature, and the form of the lambda as subsequently measured will be different from its normal form because of relaxation effects. However, such freezing-in processes are not responsible for the existence of the lambda. In glass-forming systems, on the other hand, the existence of the discontinuity at T_g is entirely due to the fact that the degree of order appropriate roughly to T_g is frozen-in at all lower temperatures, and only that part of the total order–disorder change which is appropriate to temperatures above T_g can occur in finite times.

Another common suggestion which has been made is that the fairly sharp increase in the specific heat of glass-forming systems heated through T_g is due to the appearance of a contribution from rotational

motions which are inhibited below this temperature (see, for example, Guillien 1942). It was however argued by Simon (and has recently been shown in more detail by Kauzmann) that the increase is much too great to be explained in this way, and that the rotational contribution to the specific heat just above T_g must be quite negligible. The suggestion is interesting because it has recently been set out by Prigogine and Defay (1950) in a form which is thermodynamically quite analogous to Simon's original explanation. The most important equation of our later treatment is, in fact, derived by Prigogine and Defay and assumed by them to apply to glass-forming systems as well as to the many other systems which—as we have mentioned—are thermodynamically similar. The feature common to both explanations is that a relaxation process is assumed to occur during stabilization below the normal T_g. Prigogine and Defay appear to imply that this is a structural change towards a configuration in which only reduced rotational motions are possible, which therefore has an influence upon the enthalpy. In Simon's explanation, which is our starting-point, attention is directed to the contribution of the configurational changes themselves to the value of the enthalpy. It may be noted that the relation between these two explanations is similar to that which exists between the two explanations which have commonly been put forward for the existence of a lambda transition in, say, ammonium chloride. The suggestion of Pauling (1930) was that the existence of the lambda was due to the possibility of rotational motions at higher temperatures. The now more generally accepted explanation (Lawson 1940, Alpert 1949, Nagamiya 1952) is that this again is an order–disorder transition, the changes of order being in the relative orientations of the molecules in their lattice sites.

A number of further misconceptions are briefly mentioned in later sections where appropriate to the matter in hand.

§ 3. Kinetics of the Approach to Equilibrium

In the manufacture of ordinary glass, the final process consists of a heat-treatment, usually described by the general term ' annealing ', which consists essentially of a more or less slow cooling through the region of the transformation temperature. Or, if a specimen has already been rapidly cooled, it is reheated into the same temperature region and again slowly cooled. The best known function of such treatments is to avoid the growth of—or to remove—the ' permanent ' mechanical stresses which can exist in glass if it is cooled so rapidly that a temperature gradient exists across its thickness in the neighbourhood of T_g. The removal of these mechanical stresses is in principle completely understood if we know the reactions of the glass to simple externally applied stresses, because the stresses present are in fact simple ones acting between adjacent layers. We do not discuss this further ; the flow properties of glass have been reviewed by one of us elsewhere (Jones 1948–9).

There is, however, another function of annealing which bears directly on the main subject of the present review. In the manufacture of the highest quality optical glass it is important not only that there should be no mechanical stresses (which would lead to birefringence), but also that the properties of the glass should be uniform throughout the body of the material (see Hampton 1942). Now, as we have seen, the properties of a sample of glass at ordinary temperatures are determined partly by its previous rate of cooling in the neighbourhood of T_g. It is necessary therefore to take extreme precautions to ensure that the rate of cooling through this region is the same at all points in the glass. This again requires that temperature gradients should as far as possible be absent, and that the glass be cooled as slowly as is practicable—or as may be necessary for a required degree of homogeneity. Again, if the specimen has been rapidly cooled, it may be re-heated, when changes occur of the kind which we have described by the term ' stabilization '.

Because of the practical importance, and suspected theoretical importance, of these processes, there have been very many investigations into the kinetic properties of glasses near their transformation temperatures. Unfortunately, the inorganic glasses usually studied have transformation temperatures at least as high as 500°C, where continuous observation of changes in volume or heat content would be extremely difficult, and nearly all the existing results have been obtained in indirect ways. A common approach has been to study property/temperature relationships during uniform heating at different rates ; we have already seen in fig. 3 how the effects under discussion can make their presence felt in such experiments. Another method has been to make measurements (especially of the refractive index) on samples rapidly chilled to room temperature after controlled heat-treatments near T_g. In these investigations it is assumed that the configuration corresponding to a given temperature near T_g is preserved by chilling from that temperature. For completeness, we now include a survey of the work done in this field, and of attempts made to explain the experimental results. We mention also some recent experiments, carried out by the present authors, in which direct measurements were made of the approach to equilibrium—or stabilization—in certain organic glasses having much lower transformation temperatures, and we discuss some of the implications of these results.

3.1. *Indirect Experimental Studies*

The main contributions by indirect methods have been made by A. Q. Tool and his collaborators at the U.S. Bureau of Standards. They have published about twenty papers since 1918, of which we list only the most important (Tool and Eichlin 1920, 1925, 1931 ; Tool, Lloyd and Merritt 1930 ; Tool 1946 ; Tool, Tilton and Saunders 1947). Most of the experiments were carried out on inorganic glasses based upon silica, and were of one of the following two kinds :

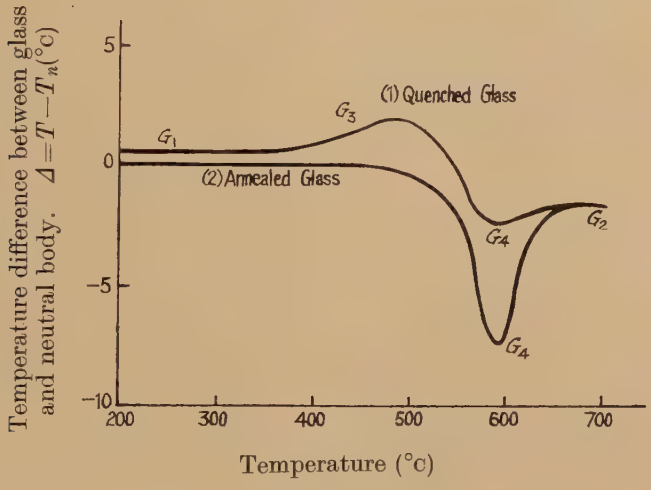
(*a*) measurements—essentially rough measurements of heat capacity—carried out by observing the differences in temperature between particles of glass and of a ' neutral body ' when heated together at uniform rate,

(*b*) measurements of the thermal expansion of glasses.

In the heating experiments the following technique was used. Glass specimens were broken into small fragments of about 1 mm diameter and then thoroughly mixed with powdered alundum. One or more differential thermocouples were arranged, each having one junction in the glass and the other in the powder. For purposes of control another couple could be attached to the wall of the furnace into which the whole material was packed. The glass could be given a known heat-treatment either before or after packing.

The experiment consisted of heating the furnace according to a controlled schedule—normally a constant rate 3–6 deg/min—and observing the temperature difference between the glass and the alundum as recorded on the thermocouples. This difference was plotted as a function of the ambient temperature. The technique was gradually improved between 1918 and 1930 so that the results became more closely reproducible. Scores of different commercial glasses were investigated.

Fig. 4



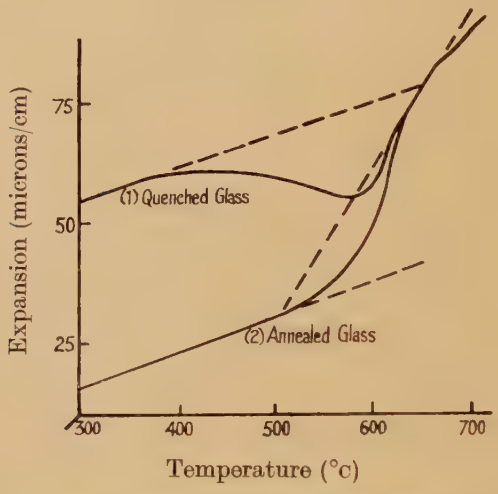
Thermal effects during heating (after Tool).

For our purpose we can reduce Tool's results to two typical curves, shown in fig. 4. If T is the temperature of the glass and T_n that of the neutral body (the alundum powder) it was found that—although the absolute value of the difference $\Delta = T - T_n$ was affected by the method of packing—reproducible results could be obtained if the ordinate scales were adjusted to make the curves coincide at the high temperatures (where the material would be in the equilibrium liquid state). In the figure, Δ is plotted against T and the curves shown apply to (1) a glass quenched from 850°C at 40 deg/min and (2) a glass annealed for several weeks at 500°C.

In general, Tool's curves showed an exothermic effect corresponding to the peak G_3 and an endothermic effect corresponding to G_4. The effect of annealing is to decrease the exothermic effect and increase the endothermic effect. As shown by curve 2, annealing at a low temperature can reduce the exothermic effect to zero but even badly annealed specimens continue to show the endothermic effect. The qualitative explanation of these results in terms of figs. 2 and 3 is immediate.

Tool obtained a number of unexpected results ; in particular he found that for Pyrex (borosilicate) glasses there were often two distinct minima of the type G_4 (Tool and Hill 1925, Saunders and Tool 1933). This seems to correspond to the observation of Douglas and Jones (1948) that in borosilicate glass of high silica content there are two relaxation processes operating at different temperatures.

Fig. 5



Volume changes during heating (after Tool).

Turning now to the experiments on thermal expansion, we illustrate a typical pair of results in fig. 5. Here, the length of the test specimen is plotted against temperature for two glasses subjected to the same treatments as applied to the samples of fig. 4 ; thus curve 1 is for a chilled specimen and curve 2 for one annealed at 500°c. The curves are again adjusted to coincide in the high temperature region. (At the highest temperatures spurious effects are shown which are due only to unavoidable deformation of the test specimen.)

In order to give a coherent account of his two types of experimental result Tool was led in 1931 to introduce the concept of ' fictive ' or ' equilibrium ' temperature. He had already observed that if the temperature of a sample of glass was held constant in the neighbourhood of T_g for a sufficient time the glass would reach a unique and permanent

state—which we recognize as that of the equilibrium liquid (or supercooled liquid). He therefore defined the fictive temperature $\bar{T}$ to be that temperature at which the glass would find itself in equilibrium if brought there sufficiently rapidly from its current state. As can be seen from fig. 5, the values of $\bar{T}$ at low temperatures for Tool's specimens 1 and 2 are 660°C and 500°C respectively. (The definition of $\bar{T}$ given here is equivalent to that introduced by us in § 4.3 in a somewhat different way.) Tool pointed out that it was unlikely that a single parameter would be adequate to give a perfect description of the configurational state of a glass. However, no experiments of this type have been sufficiently certain to demand more than one parameter. (It can be shown that the recent results of Douglas and Isard (1951) on silica can still be described by one parameter even though the expansivity of the glass is no longer independent of $\bar{T}$.)

Since in equilibrium $T-\bar{T}=0$, Tool conceived the quantity $T-\bar{T}$ as a ' physico-chemical driving force ' responsible for the changes occurring during stabilization. In his earlier publications he appears to imply that chemical reactions are in progress during stabilization in glass. Although this interpretation is perhaps not quite irrelevant in the case of complex glassy mixtures, it can only be considered generally misleading, and is certainly irrelevant if applied to single-component glasses. Tool's ideas as to the actual nature of the stabilization process and the significance of his fictive temperature have however been modified during the course of his work, and he may be said to have approached Simon's viewpoint asymptotically.

Having defined $\bar{T}$, Tool uses it to describe his experimental results. Since the expansivity is constant at high and low temperatures, it is reasonable to assume that the length of a specimen is given by

$$(l-l_0)/l_0=\alpha_l'(T-T_0)+\Delta\alpha_l(\bar{T}-T_0) \quad . \quad . \quad . \quad . \quad (3.1)$$

where α_l' is the linear expansion coefficient of the glass and $\alpha_l'+\Delta\alpha_l$ that of the liquid. This assumption will obviously break down if extended over too wide a temperature range. It will also be inadequate for the complicated glasses such as the borosilicate glasses already mentioned. Once α_l' and $\Delta\alpha_l$ are obtained from the results at high and low temperatures, eqn. (3.1) can be used to find the fictive temperature $\bar{T}$. A diagram such as fig. 5 gives the answer immediately.

An assumption analogous to that of eqn. (3.1) can be made concerning the effective specific heat (C) as measured when stabilization processes are occurring :

$$C\dot{T}=\dot{Q}=C'\dot{T}+\Delta C\dot{\bar{T}}. \quad . \quad . \quad . \quad . \quad (3.2)$$

Here C' and $C'+\Delta C$ are, respectively, the heat capacities of the glass and the liquid. It is then possible to relate the results of the two kinds of experiment. For example, the peaks and dips in the curves of fig. 4 can be related to the dimensionless quantity, $d\bar{T}/dT$—which can be

obtained from the expansion data. The thermal data indicate a minimum of $d\bar{T}/dT$ of $-1{\cdot}55$ and a maximum of 4·4. These figures compare reasonably well with $-0{\cdot}73$ and 5·4, derivable from the expansion data.

The effects of heat treatments near T_g on the properties of glasses, as measured after chilling to room temperature, have been thoroughly examined by a number of authors. The first to study this matter was Lebedeff (1926) and the method he used has since been more fully exploited by Winter (1943 a, b, 1946), MacMaster (1945), Collyer (1947), Tool, Tilton and Saunders (1947), Douglas and Jones (1948) and Douglas and Isard (1951). The property usually studied has been the refractive index.

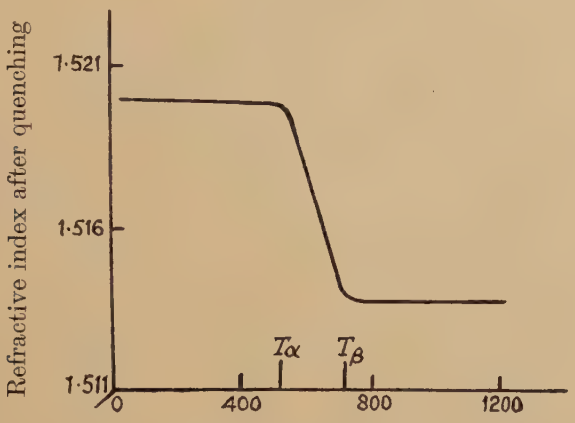
The general procedure is as follows: a large number of samples of glass are first brought to room temperature after they have all been treated alike (all quenched or all annealed). A number of specimens are then selected and placed in a constant temperature furnace. At chosen times, single specimens are removed, chilled quickly in contact with a brass block and their refractive indices measured at room temperatures. The variation of this quantity with the time of heating is found to depend both on the initial condition of the glass and on the temperature of the furnace.

Winter reports two main results: (*a*) for a given temperature of heating the refractive index as measured in this way always approaches a limit depending only on that temperature. It may approach this limit from above or below according to the initial conditions. (*b*) The approach to the limit is roughly exponential with respect to time, with a time-constant which depends on the initial condition of the glass and on the temperature of heating.

These conclusions are summarized in figs. 6 and 7. In order to explain her results Winter suggested that there was a 'transformation range' between T_α and T_β within which the glass changed its structure from a state α to a state β. She asserts that "when glass is obtained in equilibrium at T_α it possesses the maximum possible value of refractive index for a given glass. . . . At the same time it attains a molecular structure which remains constant for lower temperatures". This explanation is inconsistent with the general viewpoint which we have adopted and represents a return to ideas held by the continental school (see Mondain-Monval and Galet 1930) who claimed that the region near T_g was a fixed transition range of some sort. There is nothing in Winter's results which cannot be explained by recognizing that (*a*) with a finite rate of chilling the structure cannot be frozen-in above some temperature (T_g) and (*b*) limits upon the time-scale or the accuracy of measurement will make it impossible to reach equilibrium below some other temperature (T_α). Curves very similar to those of fig. 6 have also

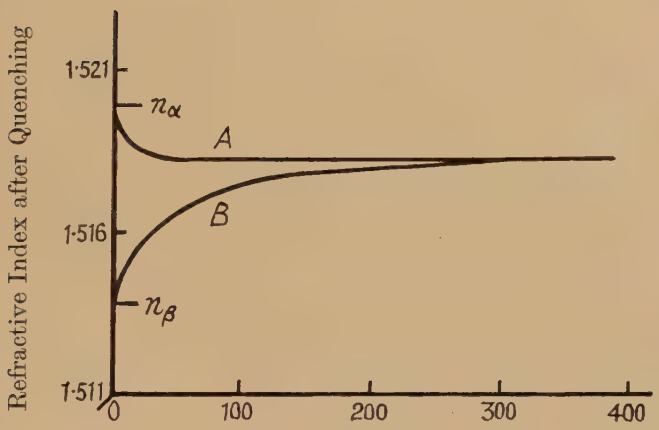
been obtained by Tool, Tilton and Saunders. However, these authors showed clearly that the flat portions at high and low temperatures must be due to limitations in the experimental technique.

Fig. 6



Refractive index of quenched samples (after Winter).

Fig. 7



The approach to equilibrium (after Winter).
Furnace temperature : 580°C ; initially stabilized at 480°C (curve A) and 720° (curve B).

Winter carried out one most interesting experiment showing the effect of pressure on the approach to equilibrium. She found that when the glass was first stabilized at T_β and then held at a test temperature T_α, the

approach to equilibrium could be accelerated by applying an external pressure of 20 kg/cm^2. She concluded also that the equilibrium value of refractive index at T_α was unaltered by the application of pressure. It seems probable that this latter conclusion is incorrect and that insufficient time had been allowed for the attainment of equilibrium.

3.2. *Direct Measurements*

Apart from some isolated and somewhat unsatisfactory results on the refractive index (Winter 1943 a, b) and viscosity (Lillie 1936) of silicate glasses, the only direct measurements of kinetic effects have been made on the organic glasses glycerol and glucose (Davies and Jones 1953). Such glasses are attractive from an experimental point of view because their transformation temperatures are at or below room temperature, and because as single component glasses they are unlikely to show the complex behaviour noted by Tool and Douglas in borosilicate glass.

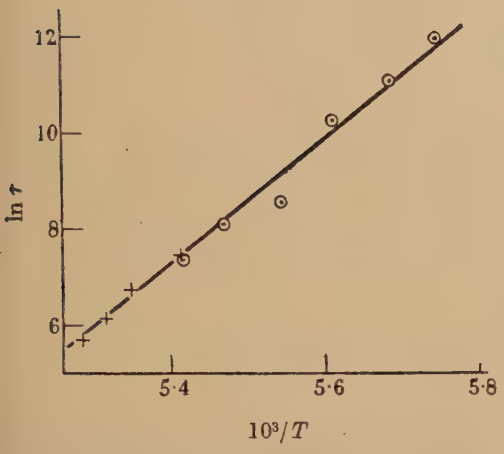
Glycerol has a transformation temperature of 180°K so that it was natural to take advantage of low-temperature calorimetric techniques and to study the thermal properties. A modified Nernst calorimeter was constructed which could be held adiabatic over long periods (up to eight hours) and in which the glass could be cooled or heated fairly rapidly when required. It was thus possible to cool or heat the specimen rapidly from an equilibrium state and then after isolating the calorimeter, to observe the change in temperature as the new equilibrium state was approached. This was equivalent to measuring the change in enthalpy with time and the approach to equilibrium could be made either from a state of higher enthalpy or a state of lower enthalpy. The experiments were not isothermal but followed lines parallel to BD′ and EF′ on the enthalpy/temperature diagram of fig. 2.

The relaxation curves found by this method showed significant departures from exponential form (with respect to time) whenever the initial enthalpy differed from the final enthalpy by more than about 0·25 cal/g. It was nevertheless possible to assign a relaxation time τ in a consistent way over the temperature range 170–190°K. (See § 3.3 for a further discussion of the detailed shape of the relaxation curves.) The collected values of $\ln \tau$ are plotted in fig. 8 against $1/T$. A reasonable line drawn through the points gives an 'activation energy', defined by $E_{\text{stab}}=R\, d \ln \tau/d\,(1/T)$, of 25 ± 2 kcal/mole.

A similar set of experiments was performed on glucose, using changes in volume instead of changes in enthalpy as a measure of the approach to equilibrium. The transformation temperature of glucose is 30°C so that in this case it was convenient to make a strictly isothermal approach to equilibrium. The curves of volume against time again showed departures from exponential shape similar to those found for glycerol. Values for τ could however still be estimated consistently. $\ln \tau$ is plotted against $1/T$ in fig. 9. The relation is again linear and the slope of the line corresponds to an activation energy, E_{stab}, of 132 ± 10 kcal/mole.

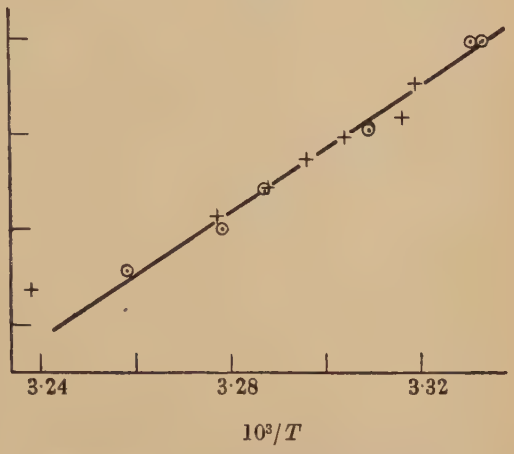
A point of some interest is that with both glucose and glycerol the departures from exponential form while approaching equilibrium differed according to the direction of approach. This is illustrated in figs. 10 and

Fig. 8



Relaxation time for stabilization of glycerol (by thermal measurements).

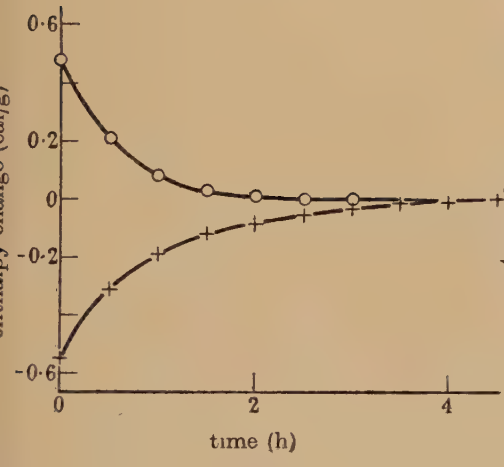
Fig. 9



Relaxation time for stabilization of glucose (by volume measurements).

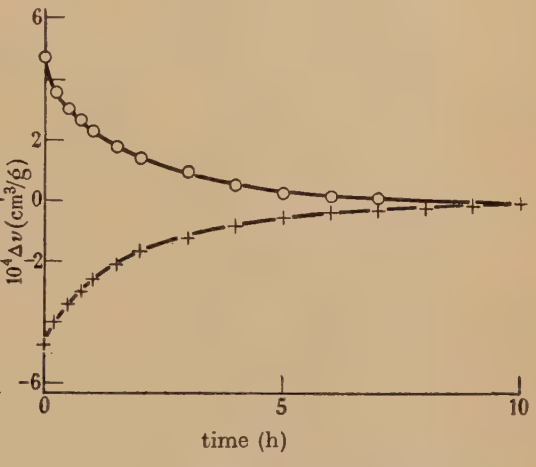
Equilibrium approached from below (+) and above (⊙).

Fig. 10



The approach to equilibrium for glycerol. Mean temperature 184·9°K.

Fig. 11



The approach to equilibrium for glucose. Temperature 304·2°K.

+ approach from below ($\bar{T}<T$) ○ approach from above ($\bar{T}>T$)

11 which show approaches to equilibrium for each substance made from opposite sides of the equilibrium state, the initial displacements being approximately equal. It will be seen that in both cases the approach from the high temperature side is the more rapid of the two.

3.3. *Analysis of the Experimental Results*

Starting from his concept of fictive temperature, Tool was able to take the first steps towards a more rational treatment of the kinetics of stabilization. He considers that the real problem is to find the rate of change of the fictive temperature and proposes two formulae :

$$d\bar{T}/dt = K(T-\bar{T}) \exp(T/k), \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3.3)$$

$$d\bar{T}/dt = K(T-\bar{T}) \exp(T/g+\bar{T}/h) \quad . \quad . \quad . \quad . \quad . \quad . \quad (3.4)$$

where K, k, g and h are constants.

The first factor involving temperature on the right hand side is the ' physico-chemical driving force ' mentioned earlier. The remaining factors are introduced to account for the viscous resistance to changes of structure. The second expression might be expected to be more significant since it allows for the possible effect of changing configuration on the viscous resistance. In neither case has he chosen a form exactly similar to any of the usual expressions for the temperature dependence of the fluidity. However, this need not invalidate the equations provided they are used only over a narrow temperature range. The form of eqn. (3.4) implies that the fictive temperature contributes to the fluidity in the same way as the true temperature. If the fluidity is taken as proportional to $\exp(E(V)/RT)$ and Tool's expression (3.1) is taken into account, then a little consideration shows that this must be at least approximately true. It is therefore not surprising to find that there have been two attempts to verify eqn. (3.4) which have been notably successful : first, Tool applied the equation in order to predict the shape of his expansion curves (§ 3.1). The agreement, although not perfect, is impressive. A more exhaustive test was made by Collyer. He was able to choose values of K, k and h so as to fit the results of measurements of refractive index (n) for quenched specimens both for fixed n and varying T and for fixed T and varying n. His work covered the range 515–550°c and at the lower end measurements were continued over periods up to 12 days.

Tool's second formula (3.4) implies that at a given temperature and for a given displacement from equilibrium the rate of approach to equilibrium will be greater when $\bar{T}>T$ than when $\bar{T}<T$. Qualitatively this can be easily understood because the volume of the glass is greater and the structure somewhat looser in the former case. This means that the potential barriers impeding structural readjustments will in turn be somewhat lower so that the ' reaction ' can proceed more quickly. Although this idea was apparently contradicted by the early results of Lillie (1936) on the viscosity of silicate glasses his results are open to objection because the initial displacements from equilibrium on either side were greatly different. The results illustrated in figs. 10 and 11 lead unequivocally to the conclusion implied by eqn. (3.4).

It is clear that before a complete explanation of the kinetics of stabilization can be given, some more realistic attempt than is represented by the above formulae must be made to represent the effect of variations

in volume upon the rate at which the micro-processes of stabilization occur at, say, a given temperature. The influence of the volume on the rate of ordinary viscous flow is known to be considerable, since below T_g, where the configurational contribution to the expansivity disappears, the calculated value of E_{visc} (equal to $d \log \eta / d(1/T)$, where η is the viscosity) is reduced by a factor of order 10 compared with its value immediately above T_g (see Jones 1948–49, Pearson 1952).

Not only is the behaviour illustrated in figs. 10 and 11 non-linear with respect to the magnitude of the displacement from equilibrium, as we have seen, but it is also non-exponential with respect to time, for large displacements. For these reasons, and because of the uncertainty we have felt in attempting to interpret reasonably even the experimental curves obtained in our direct studies, we do not venture at this stage to attempt a further systematization of the results. We must add the remark that the previous formulae we have mentioned must in reality be in the highest degree empirical. Indeed, other suggested empirical formulae (e.g. MacMaster 1945) have been found to describe the results of indirect experiments equally well. We have not mentioned them here because they were not claimed by their authors to have any physical basis, and are not illustrative of the points under discussion.

One of the most striking features of the phenomenon of stabilization is the way in which T_g can be located in terms of the value of the viscosity, and it is necessary to consider whether this can be given any quantitative interpretation. The connection was implicity recognized by Tool when he attempted to represent the kinetics of stabilization processes by a formula involving the viscosity. We have noted in § 1 a more explicit relation, namely, that T_g always occurs where the viscosity is about 10^{13} poise. For most substances this implies a Maxwell relaxation time of the order of minutes at T_g, that is, of the same order as the time taken in ordinary experimental measurements. Experiments, such as those of Cattoir and Parks (1929), on dielectric absorption give indirect support to the idea that viscosity is the controlling factor. Both the simple Debye–Einstein theory and more sophisticated rate theories (Kauzmann 1942) relate the dipole relaxation time τ_R to the viscosity through the formula $\tau_R = CV\eta/RT$ (where C is of order unity). Experiments show that for most glasses T_g is roughly fixed by finding the temperature for which τ_R is 30 minutes. The relation just quoted then suggests that η is the determining parameter.

The rough success of the idea that the viscosity is the controlling factor has veiled its inadequacy. Early indications of discrepancies were considered by Jenckel (1939). There are at least two further sets of experiments which appear to contradict this proposition directly, and certain implications of our own results weaken it further. For example, significant differences have been observed between T_g as measured experimentally in the ordinary way and as estimated in experiments on dielectric relaxation. According to the data of Kuwshinsky and Kobeko

(1938) on the dielectric constant of n-propyl alcohol, the relaxation time for dipole rotation is 30 minutes at 123±5°K. On the other hand, Parks and Huffman (1926, 1927), who measured the heat capacity of this substance, found T_g at 86–90°K. Even though the interpretation of data on dielectric relaxation is notoriously uncertain we cannot fail to take note of this very large temperature difference—which corresponds to a factor of 10^5 in viscosity. There are of course additional doubts about this comparison since the results were obtained by two sets of workers using different samples of the material.

The second set of results is a little more significant because the work was carried out systematically by one group of observers. Fox and Flory (1950) measured the volume and viscosity of a number of fractionated specimens of polystyrene. For specimens with mean molecular weights of 3000, 10 000 and 300 000 the respective transformation temperatures were 43°C, 83°C and 99°C. Below T_g the volumes of all the specimens were equal and above T_g their expansivities were equal. The values of $\log_{10}\eta$ at T_g for the specimens mentioned were, respectively, 11·31, 10·58 and 14·19, thus differing from each other by several units. These values were indeed not measured directly but were obtained by extrapolation of the results of observations taken up to 10^6 poise. Because of the comparatively rapid changes in the activation energy normally found above T_g for glass-forming liquids, the validity of this extrapolation is of course questionable, so that even this evidence is not conclusive.

However, the proposition cannot stand up to detailed examination in physical terms. Ordinary viscosity refers to shearing motion while there is no need for shearing motion to occur during stabilization. It is clear that what is required is an appropriately defined *volume* viscosity. The question of how such a quantity is to be defined is discussed in detail in § 4, and it will be shown that the results of experiments such as those illustrated in figs. 10 and 11 enable an estimate to be made of its magnitude. Further, the ratio η_v/η (where η_v is the volume viscosity) is found to differ greatly as between the two substances glycerol and glucose, so that it is hardly to be expected that η can be the decisive parameter in stabilization.

Finally, we mention an interesting point which arises from the variation of η_v with temperature, and which can be discussed even at this stage, before η_v has been exactly defined. Whatever the definition of η_v, the quantity $d\ln\tau/d(1/T)$ of § 3.2 will be expected to correspond to some kind of activation energy (E_{stab}) representative of the barrier to 'volume' viscous flow. We have already quoted values of E_{stab} for glycerol and glucose and we now assemble table 2 in which these values are compared with values of E_{visc} obtained by various authors for the same substances. It will be seen that E_{stab} and E_{visc} are obviously identical, within the accuracy of the experiments. We may therefore conclude that if the activation energy as measured is really descriptive of the barrier to some molecular rearrangement (such as the average

energy needed for a molecular rotation or migration during a ' unit flow process ') then the unit processes involved in volume flow and in shear flow are identical.

Table 2

	E_{stab}(kcal/mole)	E_{visc}(kcal/mole)
Glucose (304°K)	132 ± 10	125 ± 10 (Davies and Jones 1953) 106 (Parks, Barton, Spaght and Richardson 1943)
Glycerol (178°K)	25 ± 2	23 28 (extrapolations, using two formulae, of results of Tammann and Hesse 1926)

§ 4. Application of Thermodynamics

4.1. *The Validity of a Thermodynamic Treatment*

Before embarking on a thermodynamic discussion it is necessary to consider whether an entropy function exists for glasses. In classical thermodynamics entropy is defined by $dS = dQ_{rev}/T$ in which the subscript reminds us that it is essential for the heat flow to be measured during a reversible process. In practical calorimetry it is not necessary to worry about reversibility because the quantity actually measured is the change in energy (or enthalpy) of the system. It is assumed that the system is in equilibrium before and after the transfer of energy and that a reversible path *exists* between the initial and final states. dQ_{rev} is then taken to be the reversible heat which would be required to effect the same change as in fact occurs. For example, at constant pressure dQ_{rev} is equal to the measured quantity dH.

For glasses the situation is not quite so simple because it is not clear that a reversible path exists between, say, a given glassy state and an equilibrium liquid state. We now show that this difficulty can be overcome and discuss further the question as to whether the entropy as measured is a unique function of state.

In the case of a glass we can proceed as follows : we first stabilize the material at a relatively low temperature $\overline{T}$ below which no appreciable changes can be observed in ordinary experiments (i.e. those with time-scales of minutes or hours). Then all changes below $\overline{T}$ are reversible in a practical sense, in that the amount of irreversible change can be made as small as desired by a suitable choice of $\overline{T}$ (provided that the experimental time scale is itself of the order of minutes or hours). Thus above $\overline{T}$, changes are reversible in the classical sense, and below $\overline{T}$ they are reversible in the specified practical sense. Using this type of procedure the entropy and other thermodynamic properties of a glass of given $\overline{T}$ can be uniquely defined as functions of p and T.

It may be objected that the concept of reversibility demands the possibility of an infinite time-scale so that the practical restriction is itself contrary to thermodynamics. There are two answers to this which, taken together, are relevant to the status of all thermodynamics. In the first place, the introduction of an infinite time-scale would mean (in the case of a glass) that crystallization would certainly occur, both above and below $\bar{T}$. The discussion is thus turned to the question of the status of supercooled—and indeed of all metastable—phases. Secondly, it is well known that even the physically stable phases normally dealt with in thermodynamics may be unstable with respect to chemical changes. A familiar example of this is the inhibition of the chemical reaction in a mixture of hydrogen and oxygen at ordinary temperatures, and there is no reason why even nuclear reactions should not be considered.

Presumably we must regard *all* thermodynamic theorems as subject to certain restrictions on the types of change to be allowed. The case of glasses then ceases to be anomalous except for the fact that the characteristic restriction involved is one not usually considered in thermodynamical arguments. The restriction is that (ignoring vibrations) the positions of the molecules must remain fixed except for macroscopic changes of scale. The stabilization process which may occur when the restriction is lifted can be described as a rather unusual type of volume flow—unusual because it can occur at constant volume.

These ideas must now be given a quantitative thermodynamic formulation. It was pointed out in § 1.1 that a chemically reacting system in which the reaction can be inhibited by changing the temperature should be very similar to a glass in its thermodynamic properties. The reason why ordinary thermodynamics can give an account of such a system is that the restriction to a single phase of fixed composition can be lifted by considering the transfer of mass into the system. This makes it possible to define chemical potentials experimentally by putting the mixed phase into osmotic equilibrium with each of the pure constituents in turn. One might expect that an analogous procedure could be used to extend the formulae applicable to single phases so as to deal with glasses. In this case, however, there seems to be no simple way of defining the co-ordinate corresponding to the degree of reaction which, in the chemical case, can be determined by analysing the system or —in a more fundamental way— by allowing the various constituents to diffuse into adjacent specimens of pure substances.

In accordance with the above discussion a rather different possibility exists for dealing with glasses; namely that of bringing the glass instantaneously into the equilibrium liquid state having the same configuration. Referring to fig. 2, let us suppose that the irreversible transition along BD is interrupted at X and the temperature is quickly raised so as to bring the system to Y. Is the glass then in the equilibrium liquid state? It is a definite physical assumption that it is. As far as we know there has been no experiment designed to test this assumption directly but, on the other hand, no existing result contradicts it. (See however, § 4.6 where it

is shown that preliminary tests of certain consequences of this assumption throw doubt on its validity.) The corresponding assumption in chemical thermodynamics is that the chemical potentials (and hence the other thermodynamic variables) depend only on the instantaneous values of p, T and the concentrations. This is hardly open to doubt although it has often been explicitly recognized as an assumption. It may be worth noting that if our assumption is *false*, then when the glass is quickly brought to have the enthalpy of the equilibrium liquid the volume cannot be that of the equilibrium liquid. Or, when it is brought to have its equilibrium volume then the enthalpy cannot have its equilibrium value. Only if both properties assume the values corresponding to the equilibrium liquid at the same time is the assumption true.

Adopting the assumption, we proceed as follows : we assert that all states of a glass which can be carried into a single equilibrium liquid state by instantaneous changes of pressure and temperature can be regarded—in a certain sense—as equivalent. They can be labelled by a single value of an arbitrary ordering parameter z. It is this parameter—numerically undefined—which corresponds to the degree of reaction in chemical thermodynamics and whose properties we shall now discuss. We wish to emphasize that, by the assumption we have just made, our treatment is applicable to real systems only if their behaviour can be described in terms of a *single* ordering parameter.

4.2. *Relations between Properties of the Glass and Liquid*

In a number of rough statistical theories one or more ordering parameters such as z are introduced into a synthesized expression for the Helmholtz free energy, F, and subsequently eliminated by using the equilibrium condition $(\partial F/\partial z)_{V,T}=0$. We wish to show that it is possible to obtain useful results by the use of such a formalism without knowing the explicit expression for F. Our first approach is similar to one made by Frenkel (1946) and we begin with some purely mathematical considerations.

Let $\phi(x, y, z)$ be a differentiable function of the three independent variables x, y and z. Consider a set of surfaces defined by $u=u(x, y, z)=$constant. The relation between derivatives of ϕ at constant z and those at constant u (i.e. along the surfaces $u=$constant) are then given by formulae such as

$$\Delta(\partial\phi/\partial x)\equiv(\partial\phi/\partial x)_u-(\partial\phi/\partial x)_z=-(u_x/u_z)\phi_z \quad . \quad . \quad . \quad (4.1)$$

where we have introduced the symbol Δ for brevity. Suppose now that u is of the form $(\partial f/\partial z)_{x,y}$ where f is some ' potential ' function. With the help of eqn. (4.1) we may then derive the relations

$$\left.\begin{aligned}\Delta(\partial f_x/\partial x)&=-f_{xz}^2/f_{zz}\\ \Delta(\partial f_y/\partial y)&=-f_{yz}^2/f_{zz}\\ \Delta(\partial f_x/\partial y)=\Delta(\partial f_y/\partial x)&=-f_{xz}f_{yz}/f_{zz}\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (4.2)$$

From these it follows that

$$\Delta(\partial f_x/\partial x)\Delta(\partial f_y/\partial y)=[\Delta(\partial f_x/\partial y)]^2=[\Delta(\partial f_y/\partial x)]^2. \quad . \quad . \quad (4.3)$$

These results have an immediate application to the thermodynamics of glasses. For example, if we replace x, y, z and f by V, T, z and F respectively then the surfaces used in deriving eqns. (4.2) become $(\partial F/\partial z)_{V,T}=\text{const.}$ On the other hand the condition for thermodynamic equilibrium is $(\partial F/\partial z)_{V,T}=0$ so that the derivatives at constant u can be interpreted as properties of the equilibrium liquid. Derivatives at constant z refer to properties of the glass. As before we distinguish the latter by dashed symbols (e.g. $\alpha'=(1/V)(\partial V/\partial T)_{p,z}$ is the expansivity of the glass) while leaving the symbols referring to the equilibrium liquid affix-free ($\alpha=(1/V)(\partial V/\partial T)_{p,\,\partial F/\partial z=0}$ is the expansivity of the equilibrium liquid). In this context, the symbol Δ introduced above again signifies the difference between a property of the liquid and that of the glass ; that is, $\Delta\alpha=\alpha-\alpha'$, etc.

Equation (4.3) provides an identity connecting properties of the glass and those of the equilibrium liquid. It can take several forms. Replacing $f(x, y, z)$ by $U(V, S, z)$, $H(p, S, z)$, $G(p, T, z)$ and $F(V, T, z)$ in turn, we find

$$\left.\begin{aligned} \Delta(1/C_V)\Delta(1/\kappa_S)&=TV[\Delta(\alpha/\kappa C_V)]^2, & -\Delta(1/C_p)\Delta\kappa_S&=TV[\Delta(\alpha/C_p)]^2\\ \Delta C_p\Delta\kappa&=TV[\Delta\alpha]^2, & -\Delta C_V\Delta(1/\kappa)&=TV[\Delta(\alpha/\kappa)]^2 \end{aligned}\right\} \quad (4.4)$$

(κ_S is the adiabatic compressibility). All the equations in (4.4) can be combined to yield

$$\Delta(\kappa C_V/\alpha)\Delta(1/\alpha)=\Delta(C_p/\alpha)\Delta(\kappa/\alpha). \quad . \quad . \quad . \quad . \quad (4.5)$$

The thermodynamic properties of a substance are generally completely defined if *three* derivative properties such as C_p, α and κ are known. The identities (4.4) and (4.5) show that, provided the properties of the equilibrium liquid are known, those of the glass (in the neighbourhood of equilibrium) can be derived from *two* additional derivatives. This conclusion can be made plausible geometrically by considering that the potential surface corresponding to the equilibrium liquid must be generated by the characteristic lines of the one-parameter family of ' glass ' surfaces having constant z. The shape of a surface of constant z near a given point of contact is determined by two new independent quantities—one specifying the slope of the characteristic and another fixing the radius of curvature in a plane perpendicular to the characteristic.

An alternative way of developing formulae (4.4) and (4.5) which is completely equivalent to that already used introduces the formalism of irreversible thermodynamics. We have deliberately delayed its introduction in order to emphasize that the above formulae do not themselves rest on any theories of irreversibility. Nevertheless this method (associated with de Donder) is physically more suggestive and will be found useful at a later stage. Applications of de Donder's method in chemical thermodynamics have been comprehensively surveyed by Prigogine and Defay (1950) and these authors have also given the third of eqns. (4.4).

The main assumptions of de Donder's method can be summed up in the combined equation for the First and Second Laws :

$$dQ=TdS-Adz=dU+pdV=dH-Vdp \quad . \quad . \quad . \quad (4.6)$$

which is supposed to be true even during irreversible changes. Irreversibility is allowed for by the introduction of the term Adz, where A (the affinity) is a function of state. The rate of production of entropy (equal to the gross rate of increase of entropy in the system, dS/dt, plus the rate of increase of entropy in the surroundings $(1/T)\,dQ/dt$ is, in fact, $(A/T)\,dz/dt$, which must be positive or zero. It follows that, for a system permitting bilateral variations in z, the necessary and sufficient condition for equilibrium is $A=0$. For a stable equilibrium we require $\delta A/\delta z<0$ for all conceivable changes of state. Since

$$-A=(\partial U/\partial z)_{V,S}=(\partial H/\partial z)_{p,S}=(\partial G/\partial z)_{p,T}=(\partial F/\partial z)_{V,T},$$

it is clear that this method of treatment is virtually the same as that given earlier. However the introduction of the affinity allows the use of methods more familiar to students of thermodynamics.

We can now give, for example, an alternative derivation of the static formulae given above. The value of $\Delta\kappa$ is

$$-\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_{T,A}+\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_{T,z}=-\frac{1}{V}\left(\frac{\partial V}{\partial z}\right)_{p,T}\left(\frac{\partial z}{\partial p}\right)_{T,A}$$
$$=\frac{1}{V}\left(\frac{\partial V}{\partial z}\right)_{p,T}\left(\frac{\partial A}{\partial p}\right)_{T,z}\left(\frac{\partial z}{\partial A}\right)_{p,T}=-\frac{1}{V}\left(\frac{\partial z}{\partial A}\right)_{p,T}\left(\frac{\partial V}{\partial z}\right)^2_{p,T}.$$

In a similar way we can obtain expressions, valid at equilibrium, for all the equations of (4.2) :

$$\left.\begin{aligned}\Delta C_p&=\delta H^2/\beta T\\ \Delta\alpha&=\delta H\delta V/\beta T V\\ \Delta\kappa&=\delta V^2/\beta V\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.7)$$

where $\delta V=(\partial V/\partial z)_{p,T}$ etc. and $\beta=-(\partial A/\partial z)_{p,T}>0$—the inequality being an expression of the condition for stability. These formulae are interesting because they show immediately why ΔC_p and $\Delta\kappa$ are positive while $\Delta\alpha$ is of undetermined sign (see § 2.1). Of course δH, δV and β remain undefined numerically as long as z is not specified.

4.3. *Alternative Formulations in terms of the Fictive Temperature or Pressure*

Instead of using an undefined parameter z as we have done so far it is profitable to choose variables better adapted to particular problems. The most convenient choice for application to the results of calorimetric experiments is the fictive temperature $\overline{T}$ introduced by Tool (see § 3.1). A slight generalization of Tool's definition is suggested by the de Donder theory. Since in equilibrium the affinity is zero, we define $\overline{T}$ by the equation $A(p, \overline{T}, z)=0$ which yields $\overline{T}$ as a function of p and z. Provided the pressure is kept constant this is equivalent to Tool's definition.

In order to save the theory from being merely formal it is necessary to make the assumption that the system is near enough to equilibrium to allow us to treat derivative properties as constant. We may then write the following approximate expressions for the affinity

$$A(p,T,z)=(T-\bar{T})(\partial A/\partial T)_{p,z}=(T-\bar{T})\,\partial H/T \quad . \quad . \quad . \quad (4.8)$$

and treat the change in z as given approximately by

$$dz=-\left[\left(\frac{\partial z}{\partial \bar{A}}\right)_{p,T}\left(\frac{\partial A}{\partial \bar{T}}\right)_{p,z}d\bar{T}+\left(\frac{\partial z}{\partial \bar{A}}\right)_{p,T}\left(\frac{\partial A}{\partial \bar{p}}\right)_{T,z}dp\right]$$
$$=\frac{1}{\beta}\left[\left(\frac{\delta H}{T}\right)d\bar{T}-\delta V\,dp\right]. \quad . \quad . \quad (4.9)$$

With the help of eqn. (4.7) we may now derive

$$dQ=C_p'dT-TV\alpha dp+\Delta C_p d\bar{T} \quad . \quad . \quad . \quad . \quad (4.10)$$

and

$$dV=V\alpha' dT-V\kappa dp+V\Delta\alpha d\bar{T}. \quad . \quad . \quad . \quad . \quad (4.11)$$

It will be noted that for changes at constant pressure eqns. (4.10) and (4.11) are equivalent to eqns. (3.2) and (3.1) proposed by Tool. At the same time they show why a relaxing system may often be successfully treated by regarding it as a combination of two sub-systems (Zener 1948, Gorter 1947). For example : in considering paramagnetic relaxation $\bar{T}$ could be taken as the spin temperature and appropriate modifications made in the other variables.

Just as we have defined the fictive temperature analytically by $A(p,\bar{T},z)=0$ so we can define the fictive pressure $\bar{p}\,(T,z)$ by $A(\bar{p},T,z)=0$. The results (4.8)–(4.11) can be replaced by expressions involving $\bar{p}$ instead of $\bar{T}$. The approximations corresponding to (4.8) and (4.9) which hold near equilibrium become

$$A(p,T,z)=-(p-\bar{p})\,\delta V \quad . \quad . \quad . \quad . \quad . \quad (4.12)$$

and

$$dz=\beta^{-1}\,[(\delta H/T)dT-\delta V d\bar{p}]. \quad . \quad . \quad . \quad . \quad . \quad (4.13)$$

Comparing (4.8) and (4.12) and making use of (4.7) again, we see that the fictive temperature and fictive pressure are related by the equivalent formulae

$$\left.\begin{aligned}(T-\bar{T})\Delta\alpha&=-(p-\bar{p})\Delta\kappa\\(T-\bar{T})\Delta C_p&=-(p-\bar{p})TV\Delta\alpha\end{aligned}\right\}. \quad . \quad . \quad . \quad (4.14)$$

These show that a sudden isobaric change in temperature δT—which leaves $\bar{T}$ and p unaltered—is equivalent thermodynamically to a pressure increment of $\delta p=-(\Delta C_p'/TV\Delta\alpha)\,\delta T$. Substituting some reasonable values, $\Delta C_p=0{\cdot}2$ cal/g deg, $\Delta\alpha=2\times10^{-4}$/deg, $V=0{\cdot}7$ cm^3/g, $T=180°$K, we find that $p-\bar{p}\sim300\,(\bar{T}-T)$ atm. In the lower part of the range in which stabilization effects are observed, a value for δT of 10° would be quite easy to achieve and this gives $p-\bar{p}\sim3000$ atm. We note that the value of $\bar{p}$ is negative so that a large negative applied pressure would be necessary

to hold the system in equilibrium at the new temperature ; in the absence of such a pressure the glass contracts. We can regard $p-\bar{p}$ (a large positive quantity in this case) as providing an easily apprehended decription of Tool's ' physico-chemical driving force '. The order of magnitude of $p-\bar{p}$ makes it comparable with the internal pressures of most substances. In their molecular interpretations the two quantities must be closely linked.

For completeness we add the analogues of the approximate eqns. (4.10) and (4.11) which might be of service in interpreting the results of measurements of compressibility. They are :

$$dQ=C_p dT-TV\alpha' dp-TV\Delta\alpha d\bar{p} \quad . \quad . \quad . \quad . \quad (4.15)$$

and

$$dV=V\alpha dT-V\kappa' dp-V\Delta\kappa d\bar{p}. \quad . \quad . \quad . \quad . \quad (4.16)$$

Specific heat measurements on glasses are often used to estimate the zero-point entropy. In this connection, and for certain other purposes, it is necessary to know how to allow for the irreversible production of entropy. We have already seen that the rate of irreversible production of entropy is given by $(A/T)dz/dt$. For changes taking place either at constant pressure or at constant temperature, this expression can be given a simple graphical interpretation. At constant pressure, for example, eqns. (4.8) and (4.9) enable us to write the rate of entropy production as $(\Delta C_p(T-\bar{T})/T^2)\,(d\bar{T}/dt)$. On the other hand it is easy to see that the element of area traced out by a line on the (H, T) plane joining a (moving) point representing any glass to the point representing the corresponding equilibrium liquid is $\Delta C_p(T-\bar{T})\,d\bar{T}$. It follows that the irreversible production of entropy when the substance moves along any path on the (H, T) plane is proportional to the area swept out by the line just mentioned. In fact $\Delta S_{\text{irr}}=(\text{Area})/T^2$. The result, illustrated in fig. 12, is a generalization of a result of Bridgman (1950). A number of similar constructions are possible. Another which applies at constant pressure is obtained from the (V, T) diagram. In this case $\Delta S_{\text{irr}}=(\text{Area})\times(\Delta\alpha/T\Delta\kappa)$. The (V, p) diagram can be used for a construction at constant temperature which yields $\Delta S_{\text{irr}}=(\text{Area})/T$. This type of construction has been used to estimate the possible error in the measured zero-point entropy due to irreversible effects (Davies and Jones 1953). It was found that in a typical case (glycerol) such an error could not exceed 2% of the actual value.

4.4. *The Rate of Change with Time–Volume Viscosity*

The behaviour of a glass-forming system is determined by external and internal causes. One may suppose that the quantities dQ/dt and dV/dt are externally controlled by the experimenter. Their effects on the system are given by eqns. (4.15) and (4.16) (alternatively by eqns. (4.10) and (4.11)). Since there are three independent variables of state, one further relation is required to make the system determinate.

This must be a kinetic equation and a reasonable guess at such an equation can be made in the following way. With the help of eqns. (4.12), (4.13) and (4.7) we can write the rate of production of entropy as

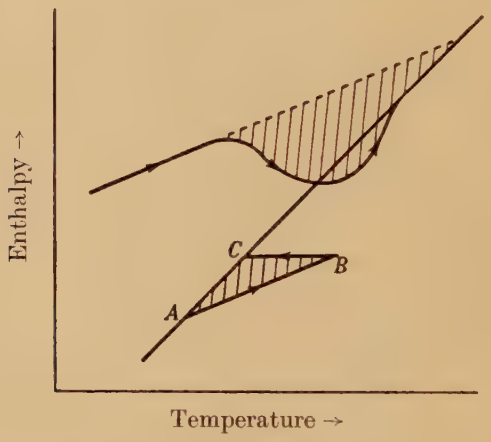
$$\left(\frac{dS}{dt}\right)_{\text{irr}} = -\frac{V(p-\bar{p})}{T}\left[\Delta\alpha\frac{dT}{dt} - \Delta\kappa\frac{d\bar{p}}{dt}\right]. \quad . \quad . \quad . \quad . \quad (4.17)$$

This can be regarded as the product of a ' force ' and a ' flow ' and as with other irreversible phenomena we may expect that sufficiently near equilibrium the flow is proportional to the force. Hence

$$\Delta\alpha\frac{dT}{dt} - \Delta\kappa\frac{d\bar{p}}{dt} = -\frac{\Delta\kappa(p-\bar{p})}{\tau} \quad . \quad . \quad . \quad . \quad . \quad (4.18)$$

where τ can be considered a relaxation time.*

Fig. 12



The irreversible production of entropy.
For the paths shown, ΔS_{irr}=(shaded area)/T^2.

The set of eqns. (4.15), (4.16) and (4.18) (or alternatively (4.10), (4.11) and the form that eqn. (4.18) takes when $\bar{p}$ is replaced by $\bar{T}$) now provides a complete description of our ' model ' glass system. It does not follow that all experiments showing relaxation actually have a relaxation time of τ—a detailed calculation must be made in each case. This can be illustrated by considering some typical experimental situations.

(1) Perhaps the simplest process of interest is pure volume relaxation at constant pressure and temperature. Equation (4.18) can then be written as $d\bar{p}/dt = -(\bar{p}-p)/\tau$ so that the relaxation time, τ_1, is equal to τ.

* The definition of τ is less arbitrary than it appears. If we define $\bar{z}$ by the equation $A(p, T, \bar{z}) = 0$, then eqn. (4.18) is easily seen to be equivalent to $dz/dt = -(z-\bar{z})/\tau$—a form which shows that τ is the natural choice for a relaxation time.

(2) Consider now a relaxation of pressure at fixed volume and temperature. By combining eqns. (4.16) and (4.18) one can show that under these conditions $dp/dt=-\kappa(p-p_0)/\kappa'\tau$. Hence the relaxation time is given by $\tau_2=\kappa'\tau/\kappa$.

(3) If the specimen is held adiabatically inside a constant-pressure container (of heat capacity C), the enthalpy of the specimen and the container is constant. Using this fact it can be seen that the temperature relaxes with a relaxation time τ_3 given by

$$\tau_3=\left(1-\frac{\Delta C_p}{C+C_p}\right)\tau.$$

The isothermal case (1) is recovered from this expression by taking the limit $C\to\infty$.

(4) For an adiabatically isolated specimen at constant pressure one can use the previous result with $C=0$. This gives $\tau_4=C_p'\tau/C_p$.

Instead of using the relaxation time τ to describe the kinetics of stabilization it is possible—as we have suggested—to introduce a suitably defined volume viscosity. In a previous publication we were led to define the volume viscosity by means of the equation $\eta_v=\tau/\Delta\kappa$. This was because the rate eqn. (4.18) can be written in the form

$$\alpha'\frac{dT}{dt}-\kappa'\frac{dp}{dt}-\frac{1}{V}\frac{dV}{dt}=\frac{(p-\bar{p})\Delta\kappa}{\tau},$$

so that for isothermal changes η_v does appear as the viscosity of a 'solid-viscous' model. For convenience we shall continue to use this definition. It is however necessary to say a few words about the confusion existing in the literature on questions connected with volume flow.

If we denote the volume strain by d, then it is easy to see from eqn. (4.18) that the formula governing *isothermal* volume flow is:

$$\kappa p+\tau\kappa'\dot{p}+d+\tau\dot{d}=0. \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.19)$$

The meaning of this relation is most easily grasped by considering the effect of a steady hydrostatic pressure applied instantaneously. The volume change consists of an instantaneous response, $-\kappa' p$, followed by an exponential relaxation to the final steady strain, $-\kappa p$. As we have already seen, this simple behaviour is only a first approximation to the behaviour of real glasses, but it is quite a good one for small displacements from equilibrium.

The first cause of confusion is that eqn. (4.19) stands in contradiction with the assumptions of classical hydrodynamics. The latter theory starts from the Poisson formula

$$p_{ij}=p_0\delta_{ij}-\lambda\dot{d}_{\alpha\alpha}\delta_{ij}-2\mu\dot{d}_{ij},$$

which for volume flow becomes

$$p-p_0=-(\lambda+(2/3)\mu)\dot{d}\equiv-\zeta\dot{d}. \quad . \quad . \quad . \quad . \quad . \quad (4.20)$$

The traditional interpretation of p_0 is that it is the pressure required to maintain the system in equilibrium at the same volume, that is, $p_0 \equiv -d/\kappa$. Equation (4.20) can therefore be written as

$$\kappa p + d + \kappa\zeta\dot{d} = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.21)$$

where ζ is a volume viscosity.* Comparing eqns. (4.19) and (4.21), we see that the latter lacks the time derivative of the pressure. It follows that for glass-forming liquids at least and possibly for other liquids (and even some gases), the classical hydrodynamical model is inappropriate. For oscillatory experiments at low frequencies ($\omega\tau \ll 1$) a certain correspondence is possible since then eqn. (4.19) becomes $\kappa p + d = 0$ in a first approximation and

$$\kappa p + d + (\tau\Delta\kappa/\kappa)\dot{d} = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.22)$$

in second approximation. Equation (4.22) can be reduced to (4.21) by writing $\zeta = \tau\Delta\kappa/\kappa^2$.

The second difficulty is due to the fact that classical hydrodynamics is a purely mechanical theory. On the other hand, the present thermodynamical treatment shows that during *adiabatic* changes the stress strain relation corresponding to (4.19) is

$$\kappa_S p + \tau_S\kappa_S'\dot{p} + d + \tau_S\dot{d} = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.23)$$

where $\tau_S = C_p'\tau/C_p$ is the 'adiabatic' relaxation time appearing in case (4) above. It follows that any definition of volume viscosity depending on the form of the flow eqns. (4.19) and (4.22) or (4.23) must have different numerical values according as the flow is isothermal or adiabatic.

It is clear that the difficulties indicated here are due to deficiencies in both theory and experiment. Until an agreed nomenclature had been reached it might have been better to have refrained from using the term 'volume viscosity'. Our definition is, however, unambiguous and the numerical values given later can if necessary be adapted to other definitions.

4.5. *Applications to other Systems*

As examples of the way in which the formulae developed above can be applied to systems other than glass-forming systems we discuss briefly their application to chemical reactions, order–disorder transitions and the absorption of sound in fluids.

For chemically reacting systems we can define z explicitly as the degree of reaction which, for the reaction $\Sigma_i \nu_i M_i = 0$, is related to the mole numbers (N_i) and stochiometric coefficients (ν_i) by $dN_i = \nu_i\, dz$. The affinity is then $A = -\delta G = -\Sigma_i \nu_i g_i$ where the g_i refer to partial molar Gibbs free energies or chemical potentials.

The significance of the symbols introduced previously is now as follows: the undashed derivative symbols refer to a system in which the reaction is allowed to proceed, while dashed properties refer to the same system when the reaction is inhibited in some way. It is immaterial

* The volume viscosity is sometimes defined as λ instead of ζ.

whether the inhibition is due to lack of a catalyst or to the fact that physical stimuli are being applied which are too rapid to be followed by the reaction.

The three formulae in (4.7) are now of enhanced significance because, in contrast to the situation with glass-forming substances, all the quantities on the right-hand side can in principle be measured independently : δH is simply the heat of reaction and δV the volume change of the reaction. β cannot be expressed quite so simply but it can be related to other properties of the solution in a number of ways. A typical formula is

$$\beta=\sum_{i,j}\nu_i\nu_j\,\partial g_i/\partial N_j. \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.24)$$

Equation (4.24) can be simplified in the case of ideal solutions and becomes

$$\beta=\frac{RT}{2N}\sum_{i,j}N_iN_j\left(\frac{\nu_i}{N_i}-\frac{\nu_j}{N_j}\right)^2 \quad . \quad . \quad . \quad . \quad . \quad (4.25)$$

—an expression which depends only on the concentrations and stochiometric coefficients. This result is, however, somewhat deceptive because in order to find the equilibrium concentrations in any particular case the reaction parameters must be known. The final stage of the reckoning is usually more or less complicated. A simple case is worth noting because the answer is in this context perhaps rather suprising. Consider an isomeric reaction in an ideal gas represented by the equation $-M_1+M_2=0$, i.e. $\nu_1=-1$, $\nu_2=1$. In virtue of the nature of the solution one finds that $\delta V=0$ and $\delta H=\delta U$. If eqn. (4.25) is evaluated with the correct equilibrium concentrations it is found that

$$\Delta C_p=NR(\delta U/2RT)^2/(\cosh(\delta U/2RT))^2 \quad . \quad . \quad . \quad (4.26)$$

while of course $\Delta\alpha=0$ and $\Delta\kappa=0$. The formula appearing in (4.26) is the Schottky formula and is well known in connection with specific heat ' anomalies '. The usual derivation is from a partition function with a single excited energy level. The fact that $\Delta\kappa=0$ implies that under isothermal compression there are no relaxation effects. This is not true for adiabatic compression. The physical explanation of this pair of facts is simply that the equilibrium degree of reaction depends only on the temperature.

The discussion given here could clearly be extended so as to include more than one chemical reaction. In the same way, the treatment of glasses could be modified to allow for more than one type of relaxation process. There are two general results of such an extension which are worth noting.

The first is that the equalities of eqn. (4.4) become inequalities with the left hand side greater than the right hand side. (Equation (4.5) also becomes an inequality but of undetermined sense.) This can be quickly seen with the help of a slightly more elaborate notation. If there are n independent ordering variables z^i the fundamental de Donder eqn. (4.6) becomes

$$dU=TdS-pdV-\Sigma_iA_i\,dz^i. \quad . \quad . \quad . \quad . \quad . \quad (4.27)$$

The conditions for equilibrium are $A_i=0$ $(i=1, 2, \ldots n)$. We introduce the generalized abbreviations :

$$\delta V_i=(\partial V/\partial z^i)_{p,T}, \quad \delta H_i=(\partial H/\partial z^i)_{p,T}, \quad \beta_{ij}=\beta_{ji}=-\partial A_i/\partial z^j, \quad \beta^{ij}=\beta_{ij}^{-1}$$

and, as before, distinguish the properties of the glass (with *all* z^j fixed) by a dash while leaving the equilibrium properties (with *all* $A_j=0$) affix-free. The generalizations of eqns. (4.7) are then seen to be

$$\left.\begin{aligned} \Delta C_p = C_p - C_p' &= \sum_{i,j} \beta^{ij}\,\delta H_i\,\delta H_j/T \\ \Delta\alpha = \alpha - \alpha' &= \sum_{i,j} \beta^{ij}\,\delta H_i\,\delta V_j/TV \\ \Delta\kappa = \kappa - \kappa' &= \sum_{i,j} \beta^{ij}\,\delta V_i\,\delta V_j/V \end{aligned}\right\} \quad . \quad . \quad . \quad (4.28)$$

The order variables, z^i, can be subjected to an arbitrary linear transformation. The transformation of the A_j must then be such that $\Sigma_i A_i dz^i$ remains invariant in form and magnitude. Under these conditions, the symmetric matrix β^{ij} can be reduced to diagonal form. The Schwarz inequality applied to the right hand side of eqn. (4.28) then yields

$$\Delta\kappa\,\Delta C_p \geqslant TV\,\Delta\alpha^2. \quad . \quad . \quad . \quad . \quad . \quad . \quad (4.29)$$

This is the desired generalization of the third member of eqn. (4.4). The others can be treated similarly. It may be observed that equalities can be recovered not only when there is a single relaxation parameter but also if it should happen that the heats of reaction are proportional to the volume changes of the reactions, i.e. $\delta H_i = k\delta V_i$ $(i=1, 2, \ldots n)$ where k is a constant. There is no *a priori* reason to expect this proportionality.

The second consequence of having more than one ordering parameter will merely be stated. It turns out that for n parameters the ' kinetic equation of state '—eqn. (4.19) is the prototype—contains time derivatives of pressure and strain up to the nth order. It therefore provides for non-exponential decay curves.

Another possible application of the formalism given here is to order-disorder transitions and, indeed, to lambda transitions in general. We have already indicated (see § 2.1) that the specific heat which would be measured if the disordering process were inhibited (C_p') can be reasonably estimated by drawing a smooth curve between parts of the experimental curve at high and low temperature—where z does not change with temperature. ΔC_p is then just the excess specific heat above this smooth curve ; $\Delta\alpha$ and $\Delta\kappa$ can be found in the same way. If experimental values are substituted in the static relations (4.4), it should be possible to test whether a single ordering parameter is sufficient to describe the transition or whether—in view of the above discussion—more than one such parameter is needed. This question will be examined in the next section.

Sound absorption in fluids provides a final example of a field of study in which the thermodynamics of relaxation can be usefully applied. It can be shown, that if shearing effects are neglected and the frequency is high enough for the motion to be adiabatic, then the propagation of sound in a medium with a single ordering parameter is controlled by two independent quantities : (*a*) the adiabatic relaxation time, τ_S, whose inverse gives the frequency of maximum absorption ; (*b*) the quantity $\Delta\kappa_S/\kappa_S$ which gives a measure of the maximum absorption per wavelength and also of the fractional increment in velocity as the frequency is raised

through the absorption region. The thermodynamic theory gives no additional information about τ_S which—as in the case of the stabilization of glass—must be measured or estimated independently. However the static formulae (4.4) and the ' chemical ' results (4.7) do provide useful information about $\Delta\kappa_S$.

If the relaxation is due to a chemical reaction, then $\Delta\kappa_S$ can be expressed in terms of the reaction parameters in a number of ways. A characteristic formula is

$$\Delta\kappa_S = \frac{VC_p}{\beta C_p'}\left(\frac{\delta V}{V} - \frac{\alpha\delta H}{C_p}\right)^2 \quad . \quad . \quad . \quad . \quad . \quad (4.30)$$

which illustrates how the quantity β (whose evaluation was mentioned above) enters into the result. A number of simple processes have been treated in detail, including dissociation and isomeric changes in ideal solution. For the case of gas reactions the results agree with those obtained by kinetic arguments (Einstein 1920, Herzfeld and Rice 1928, Kneser 1931). Equation (4.30) illustrates the point made earlier: $\Delta\kappa_S$ can be different from zero even if $\delta V = 0$ and there is no isothermal relaxation.

For structural relaxation, $\Delta\kappa_S$ can be expressed in terms of the more easily measured quantities, ΔC_p and $\Delta\alpha$:

$$\Delta\kappa_S = TV\frac{C_p}{C_p'}\frac{(\Delta\alpha - \alpha\Delta C_p/C_p)^2}{\Delta C_p} . \quad . \quad . \quad . \quad . \quad (4.31)$$

Unfortunately this equation cannot be used to provide an independent prediction of the sound absorption using results of measurements on the glass and the liquid near T_g. This is because shear effects also become important just at the frequency ($\sim 1/\tau_S$) of maximum structural absorption. The simple theory of propagation is then no longer applicable. (For further discussion of the topics of this section, see Prigogine and Defay 1950, Davies 1953.)

4.6. *Comparison with Experiment*

We are now in a position to assess the experimental facts presented in §§ 2 and 3 in the light of the thermodynamic discussion of the present section. There are two main tasks. In the first place the static relations of eqn. (4.4) make firm predictions about the data of § 2, provided that the process of glass formation can be described in terms of one parameter. Secondly, the kinetic parameter τ (or alternatively the volume viscosity) of § 4.4 can be related to the results of experiments on stabilization.

The test of eqn. (4.4) is summarized in table 3. The figures in the last column, except those for colophonium (which are probably the least reliable), are greater than unity. According to § 4.5 this implies that more than one parameter is required to deal with these substances. It is necessary to make reservations about the reliability of the values given in this table. For example, values of $\Delta\alpha$ and $\Delta\kappa$ for rubber refer to a vulcanized specimen containing 25% of sulphur, while the value of ΔC_p is that of a pure specimen (at a much lower temperature). Data for polystyrene refer to specimens which differ greatly in mean molecular

weight. Nevertheless the conclusion must be that, for glasses, the one-parameter model is not adequate. Improved thermal and kinetic measurements on a suitable *single* specimen may point the way to an improvement in the theory.

Table 3

(c.g.s. units)	T_g	V	$10^4\Delta\alpha$	$10^{-6}\Delta C_p$	$10^{12}\Delta\kappa$	$\frac{\Delta\kappa\Delta C_p}{TV\Delta\kappa^2}$
Colophonium	300	0·93	3 ca.	0·55	10 ca.	0·2 ca.
Selenium	300	0·24	2·5	1·9	5·8	2·4
Glucose	300	0·66	2·6	7·7	6·1	3·7
Rubber	320	1·1	3·1	6·0	37	8·3
Polystyrene	350	1·0	2·0	2·9	75	16

As a further—and rather less satisfactory—test of the static formulae we have analysed the data of Lawson (1940) and Simon, von Simson and Ruhemann (1927) on the lambda transition of ammonium chloride. The appropriate definitions of ΔC_p, etc., have already been given. In this case it is convenient to use another member of eqn. (4.4). The results are shown in table 4. Only above the transition temperature is the departure from unity of the second row in the sense demanded by the inequalities of § 4.5. In this region Lawson was troubled by hysteresis effects. Below the transition temperature a slight discrepancy in the temperature scales of the two sets of experiments would greatly affect the numerical values. It is therefore just possible in this case that better experimental material would improve agreement with eqn. (4.4).

Table 4

T(°K)											
230	232	234	236	238	240	246	248	250	255	260	265
$-TV\left\{\Delta\left(\frac{\alpha}{C_p}\right)\right\}^2\Big/\Delta\left(\frac{1}{C_p}\right)\Delta\kappa_S$											
6·0	5·6	3·7	2·0	1·9	1·4	0·21	0·22	0·20	0·24	0·38	0·50

We conclude by using the kinetic theory of § 4.4 to restate the results of § 3 in terms of the volume viscosity. The first step is to translate the observed relaxation time into the quantity τ which is independent of the method of observation. This is done in the way indicated by the examples (1)–(4) of § 4.4. The volume viscosity is then obtained by means of $\eta_v = \tau/\Delta\kappa$. The results are shown in table 5, together with estimates of η (see Davies and Jones 1953). $\Delta\kappa$ has not been measured for glycerol so for consistency we have used the value $TV\Delta\alpha^2/\Delta C_p$ in both cases. This makes the cited values of η_v and of η_v/η too large by a factor of 3·7 (see above).

Table 5

	$10^{13}\Delta\kappa$	$10^{-4}\tau$	$10^{-15}\eta_v$	$10^{-15}\eta$	η_v/η
Glucose (304°K)	16·4	1·01	4·5	0·024	190
Glycerol (178°K)	9·9	2·3	31	2·5 (two estimates)	12
				6·0 (two estimates)	5

While, as shown in § 4.4, the actual definition of the volume viscosity is to a certain extent arbitrary, it is clear that (see § 3.3) the numerical value of the ratio η_v/η can differ markedly from unity and also vary from one substance to another. This agrees with the observations of Liebermann (1949) made on ' ordinary ' liquids. Also, we have seen that η_v/η is independent of temperature because the activation energies E_{visc} and E_{stab} are equal.

One of us (R. O. D.) wishes to acknowledge the award by the University of London of an I.C.I. Fellowship.

References

Alfrey, T., 1948, *Mechanical Properties of High Polymers* (New York : Interscience).

Alpert, N. L., 1949, *Phys. Rev.*, **75,** 398.

Bernal, J. D., 1937, *Trans. Faraday Soc.*, **33,** 27.

Bernal, J. D., and Fowler, R. H., 1933, *J. Chem. Phys.*, **1,** 515.

Borelius, G., and Paulson, K. A., 1946, *Ark. Mat. Astr. Fys.*, **33**A, N : o 7.

Boyer, R. F., and Spencer, R. S., 1946, *J. Appl. Phys.*, **17,** 398.

Bridgman, P. W., 1950, *Rev. Mod. Phys.*, **22,** 56.

Buchdahl, R., and Nielsen, L. E., 1950, *J. Appl. Phys.*, **21,** 482.

Cattoir, F. R., and Parks, G. S., 1929, *J. Phys. Chem.*, **33,** 879.

Collyer, P. W., 1947, *J. Amer. Cer. Soc.*, **30,** 338.

Davies, R. O., 1953 (to be published).

Davies, R. O., and Jones, G. O., 1953, *Proc. Roy. Soc.* A, **217,** 26.

Douglas, R. W., and Isard, J. O., 1951, *J. Soc. Glass Techn.* **35,** 206.

Douglas, R. W., and Jones, G. A., 1948, *J. Soc. Glass Techn.*, **32,** 309.

Dugdale, J. S., and Simon, F. E., 1951, *Proceedings of the International Conference on Low Temperature Physics* (Oxford).

Einstein, A., 1920, *Sitz. Ber. preuss. Akad. Wiss.*, 380.

Elsasser, W. M., 1950, *Rev. Mod. Phys.*, **22,** 1.

Fox, T. G., and Flory, P. J., 1950, *J. Appl. Phys.*, **21,** 581.

Frenkel, J., 1935, *Acta Physicochim., U.S.S.R.*, **3,** 913 ; 1937, *Trans. Faraday Soc.*, **33,** 58 ; 1946, *Kinetic Theory of Liquids*, p. 208 (Oxford).

Gee, G., 1947, *Quart. Rev.*, **1,** 265.

Gibson, G. E., and Giauque, W. F., 1923, *J. Amer. Chem. Soc.*, **45,** 93.

Gingrich, N. S., 1943, *Rev. Mod. Phys.*, **15,** 90.

Gorter, G. J., 1947, *Paramagnetic Relaxation* (Amsterdam : Elsevier).

Guillien, R., 1942, *Comptes Rendus Acad. Sci., Paris*, **214,** 820.

Hampton, W., 1942, *Proc. Phys. Soc.*, **54,** 391.

Haseda, T., Ôtsubo, A., and Kanda, E., 1950, *Sci. Rep. Res. Inst. Tôhoku Univ.*, Ser. A, **2,** 16.

Herzfeld, K. F., and Rice, F. O., 1928, *Phys. Rev.*, **31,** 691.

Holland, F. A., Huggill, J. A. W., and Jones, G. O., 1951, *Proc. Roy. Soc.* A, **207,** 268.

Jenckel, E., 1939, *Z. Elektrochem.*, **45,** 202.

Jones, G. O., 1948–9, *Reports on Progress in Physics*, **12,** 133 (London : The Physical Society).

Jones, G. O., and Simon, F. E., 1949, *Endeavour*, **8,** 175.

Kauzmann, W., 1942, *Rev. Mod. Phys.*, **14,** 12 ; 1948, *Chem. Rev.*, **43,** 219.

Kuwshinsky, E., and Kobeko, P., 1938, *J. Tech. Phys., U.S.S.R.*, **5,** 401.

Kneser, H. O., 1931, *Ann. Phys., Leipzig*, 5, **11,** 761, 763.

Landau, L., and Lifshitz, E., 1938, *Statistical Physics* (Oxford: University Press), p. 200.

Lawson, A. W., 1940, *Phys. Rev.*, **57,** 417.
Lebedeff, A. A., 1926, *Rev. Opt. (Théor. Instrum.)*, **5,** 1.
Lennard-Jones, J. E., and Devonshire, A. F., 1937, *Proc. Roy. Soc.* A, **163,** 53 ; 1938, *Ibid.*, **165,** 1.
Lewis, G. N., and Gibson, G. E., 1920, *J. Amer. Chem. Soc.*, **42,** 1529.
Liebermann, L. N., 1949, *Phys. Rev.*, **75,** 1415.
Lillie, H. R., 1936, *J. Amer. Cer. Soc.*, **19,** 45.
MacMaster, H. A., 1945, *J. Amer. Cer. Soc.*, **28,** 1.
Mondain-Monval, P., and Galet, P., 1930, *C.R. Acad. Sci., Paris*, **190,** 120.
Mott, N. F., and Gurney, R. W., 1939, *Trans. Faraday Soc.*, **35,** 364.
Nagamiya, T., 1952, *Changements de Phase* (Paris : Société de Chimie Physique), p. 251.
Oblad, A. G., and Newton, R. F., 1937, *J. Amer. Chem. Soc.*, **59,** 2495.
Parks, G. S., Barton, L. E., Spaght, M. E., and Richardson, J. W., 1934, *Physics*, **5,** 193.
Parks, G. S., and Huffman, H. M., 1926, *J. Amer. Chem. Soc.*, **48,** 2788 ; 1927, *J. Phys. Chem.*, **31,** 1842.
Pauling, L., 1930, *Phys. Rev.*, **36,** 430.
Pauling, L., and Tolman, R. C., 1925, *J. Amer. Chem. Soc.*, **47,** 2148.
Pearson, S., 1952, *J. Soc. Glass Techn.*, **36,** 105.
Pople, J. A., 1953, *Faraday Society Discussion on Solutions* (to be published).
Prigogine, I., and Defay, R., 1950, *Thermodynamique Chimique*, new ed. (Liège : Édition Desoer).
Pryde, J. A., and Jones, G. O., 1952, *Nature, Lond.*, **170,** 685.
Saunders, J. B., and Tool, A. Q., 1933, *J. Res. Nat. Bur. Stand.*, **11,** 799.
Scott, A. H., 1935, *J. Res. Nat. Bur. Stand.*, **14,** 99.
Simon, F. E., 1930, *Ergebn. exakt. Naturwiss.*, **9,** 244.
Simon, F. E., and Lange, F., 1926, *Z. Phys.*, **38,** 227.
Simon, F. E., von Simson, C., and Ruhemann, M., 1927, *Z. Phys. Chem.* A, **129,** 339.
Staronka, L., 1939, *Roczn. Chem.*, **19,** 201.
Staveley, L. A. K., Hart, K. R., and Tupman, W. I., 1953, *Faraday Society Discussion on Solutions* (to be published).
Tammann, G., 1933, *Der Glaszustand* (Leipzig : Voss), p. 18.
Tammann, G., and Hesse, W., 1926, *Z. anorg. allg. Chem.*, **156,** 245.
Tammann, G., and Jellinghaus, W., 1929, *Ann. Phys., Leipzig*, 5, **3,** 264.
Tool, A. Q., 1946, *J. Res. Nat. Bur. Stand.*, **37,** 73 ; reprinted 1946, *J. Amer. Cer. Soc.*, **29,** 240.
Tool, A. Q., and Eichlin, C. G., 1920, *J. Opt. Soc. Amer.*, **4,** 340 ; 1925, *J. Amer. Cer. Soc.*, **8,** 1 ; 1931, *Ibid.*, **14,** 276.
Tool, A. Q., and Hill, E. E., 1925, *J. Soc. Glass Techn.*, **9,** 185.
Tool, A. Q., Lloyd, D. B., and Merritt, G. E., 1930, *J. Amer Cer. Soc.*, **13,** 632.
Tool, A. Q., Tilton, L. W., and Saunders, J. B., 1947, *J. Res. Nat. Bur. Stand.*, **38,** 519.
Warren, B. E., 1940, *Chem. Rev.*, **26,** 237.
White, W. P., 1919, *Amer. J. Sci.*, **47,** 1.
Wietzel, R., 1921, *Z. anorg. Chem.*, **116,** 71.
Winter, A., 1943 a, *J. Amer. Cer. Soc.*, **26,** 189 ; 1943 b, *Ibid.*, **26,** 277 ; 1946, *Bull. Inst. Verre, Mars*, p. 3.
Zener, C., 1948, *Elasticity and Anelasticity of Metals* (Chicago : University Press).

We beg to acknowledge the source of the following text-figures :
Figures 2, 8, 9, 10, 11 and 12 from Davies and Jones, 1953, *Proc. Roy. Soc.* A, **217,** 26–42.

The Scientific Work of René Descartes

(1596—1650)

By

J. F. SCOTT, B.A., M.Sc., Ph.D

With a foreword by H. W. TURNBULL, M.A., F.R.S.

This book puts the chief mathematical and physical discoveries of Descartes in an accessible form, and fills an outstanding gap upon the shelf devoted to the history of philosophy and science.

There is to be found in this volume the considerable contribution that Descartes made to the physical sciences, which involved much accurate work in geometrical optics and its bearing upon the practical problem of fashioning lenses, as also the deeper problems of light and sight and colour. The careful treatment that Dr. Scott has accorded to this work of Descartes is welcome, is well worth reading and will be an asset to all libraries. Publication is recommended and approved by the Publication Fund Committee of the University of London.

212 *pages*, 7″ × 10″, *amply illustrated.*

Price £1 - 0 - 0 net **Published July 1952**

Printed & Published by

TAYLOR & FRANCIS, LTD.

RED LION COURT, FLEET STREET, LONDON, E.C.4

Atomic Scientists' News

Journal of the Atomic Scientists' Association

- The Atomic Scientists' Association is an association of scientists whose work has given them special knowledge of the consequences for the world of the use and misuse of atomic energy. To make known the true facts about atomic energy and its implications, it publishes every two months the **Atomic Scientists' News.**

- Full membership of the Association is open to all scientists able to put before the public an informed opinion upon some aspect of atomic energy. Others interested in its work may become associate members. Enquiries c/o Professor J. Rotblat, Physics Department, St. Bartholomew's Hospital, London, E.C.1.

Vol. 2 **MAY 1953** **No. 5**

CONTENTS

Price to non-members 6/– (*plus postage*) **Annual subscription £1 12s. 6d.** (*post free*)

Non-members apply to the Printers and Publishers:—

Messrs. TAYLOR & FRANCIS, LTD.
Red Lion Court, Fleet Street, London, E.C.4